Conducting Usability Testing in Agile Software Development Environment

University of Oulu
Department of Information Processing Science
Master's Thesis
Sakari Järvelä
5.6.2021

Abstract

This Master's thesis conducted a qualitative approached case study with triangulation to action research for usability testing in Agile software development. First, there was done literature review to create theoretical basis for research. After literature review, the research method was constructed from usability testing methods and qualitative research tools. When research method was explained, the implementation of the research method was opened to tell how the created research method was used. Explained implementation was followed by presenting the findings. Then discussion conducted cohesion between literature, research method, implementation, and findings. Finally, conclusions summarized the main points of the thesis.

Literature review brought up several usability testing methods. Research method formed to use UCD, interview, SUS, and TAP from the usability testing methods that were found. Implementation conducted eight questions for semi structured interview, five tasks for second sprint's TAP and four tasks for third sprint's SUS.

It was identified that TAP and SUS can be used in the same usability testing session. Moreover, there were found nine hypotheses to be tested. Eight of those hypotheses were successfully tested. It was also found that action research requires adaption during the research and Agile generates changes in the plans successfully.

Usability testing should be done by using at least one usability testing method and taking end user into test sessions as soon as possible. Semi structured interview's benefit is to have more open conversation type moment with the test user. TAP achieves descriptive analysis considering how test user experiences usability of the application. SUS offers general usability level of the application in the numeric scale. Improving steps for usability testing can be seen from type of the need for the different usability responses. Another way to identify improvement is to test a new usability testing method. One way to bring usability testing in Agile software development is to conduct usability testing at the end of every sprint.

Keywords
Usability testing, Agile, AUCDI, qualitative, case study, action research, mixed methods, triangulation, SUS, TAP, interview, UCD

Supervisor
Ph.D. Information Processing Science, University Lecturer Mikko Rajanen

Foreword

My journey of master's studies is at its end. The support of family and scientific community has been outstanding. University of Oulu offered comprehensive education and tools to prepare for this final piece of studies. I hope the readers of this thesis enjoy the outcome. Hopefully, readers will be motivated to push through all the obstacles.

Thank you for reaching this far.

Sakari Järvelä

Ii, June 6, 2021

Contents

1. Introduction

This Master's Thesis presents a research to identify suitable ways of conducting usability testing in Agile software development environment. The research is carried out for practical and scientific matter. Usability testing is essential to improve user experience of a software product. The timing of the usability testing in the software development process is critical. Finding the most optimal phase for the actual testing is important. *Sovelluskehittäjät ryhmä* offered a practical environment for this research of usability testing process.

Agile software development has become a markable method that is used widely in software development. In contrast, usability of an application is a strategic pinpoint in competition of ICT. From these two methods both have scientific research material, but the integration of these methods has not been that much researched even when Agile has become popular software development method. Recent scientific circumstance between Agile and usability testing has been that the literature material about the topic barely exists. Therefore, the topic itself is important in the time this research was done.

One of the most important aspects in usability testing is to test the application with actual end users of the application. Agile itself offers layer to operate closely with the customer when taking the end user in testing sessions is an operation model that usability testing is created for. This explains the reason of this thesis examining the adequacy of different usability testing methods in Agile development context.

In application development, usability is a central way to indicate the quality of the application (Rajanen & Iivari, 2007, p. 1). Moreover, usability's price and advantages can be evaluated through the usability examination of models that define them, respectively. There exists a lot of research conducted for the usability testing of the web sites (Bordac & Rainwater, 2008, p. 115). It was pointed out that Jakob Nielsen, Jared Spool, Joshua Porter and Steve Krug presented creations to support the idea of the usability testing. The Research outcome was that the improvement of product and method can be assisted by usability testing (Winter, Rönkkö & Hellman, 2009, p. 115). The amount of research and valuable improvements support the logic to find vital information about the usability testing. This case study around the usability testing topic was still justified because straight implementation suggestions were missing in the researched literature.

The design problems the user meets are identified by usability testing. The Usability testing is compulsory when developing a software for the security-crucial systems or the government projects. Organizations and users both benefit when participating in the usability testing. The design issues can be repaired after locating them, when using usability testing that can be done without real functionality of the system. The essential domain expertise and the IT background are essential for the end-user in usability testing. (Bandi & Heeler, 2013, p. 1.)

Usability testing is a part of User Centered Design (UCD) (Salah, Paige & Cairns, 2014, pp. 100, 103). Salah, Paige, and Cairns (2014) revealed that Agile and the UCD communities have been interested in the Agile and User Centered Design Integration (AUCDI) (see Fig 1.). The combination of UCD and Agile operated as one point of view for this study to adapt usability testing.

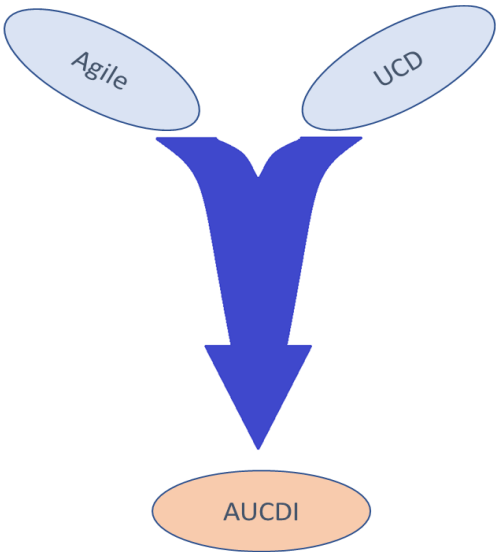

Figure 1. AUCDI as a result of combining Agile and UCD. Sakari Järvelä, 2021.

Before releasing an application, valuable information about usability issues in it can be discovered via traditional lab-based usability testing, even though it cannot secure the possible issues when the application is used in various conditions (Ferre, Villalba, Julio & Zhu, 2017, p. 115). When users run an application, the usability testing measures the usefulness and expediency of the application during the usage (Rahayu et al., 2016, p. 1). For measuring, one can gain feedback during the development process with usability testing.

This thesis was responding to the research questions that were stated as follows:

1. How should the usability testing be done in agile development context?
2. What are the benefits reached with specific usability testing methods in agile development context?
3. How could the usability testing process be improved in agile development context?
4. How could the usability testing be brought in Agile and SCRUM environments?

The Main operative research method for this research is a case study with a qualitative emphasis. First, a literature review about usability testing was performed. The goal was specially to propose with improvements to the practical process through scientific material that was found. To fulfil the research method, an action research is a concept that is applied in this research. Moreover, the literature material contained methods to define the usability level of a software product. These methods, defined by scientists, were used as research tools.

The main contribution of this case study was to implement usability testing method set in Agile software development. Qualitative research method types were interview and usage of TAP. Quantitative type of approach was conducted by using SUS.

The Structure of this Master´s Thesis was formed by starting with the Introduction that is followed by the Literature Review. The third chapter was made to explain the Research Method used for the Implementation of the formed research method. Next, the Findings

were presented and then the Literature Research and Findings were summarized in the Discussion chapter. The Final chapter consists of the conclusions of the overall research.

2. Prior Research

In this chapter the theoretical background is formed to contribute to the case-specific research method in the next chapter. Firstly, the backgrounds of the chosen material are explained to gain understanding of the usability testing contexts of different authors. The rest of the sub chapters contain literature sections that are related to the research questions of this exact research. The reason for the research-question-specific literature reviews is to operate as a scientific proof to justify actions taken in the implementation and research method definition phases. The dependence over context with the methods is possible, but not implicit (Losada, 2018, p. 7).

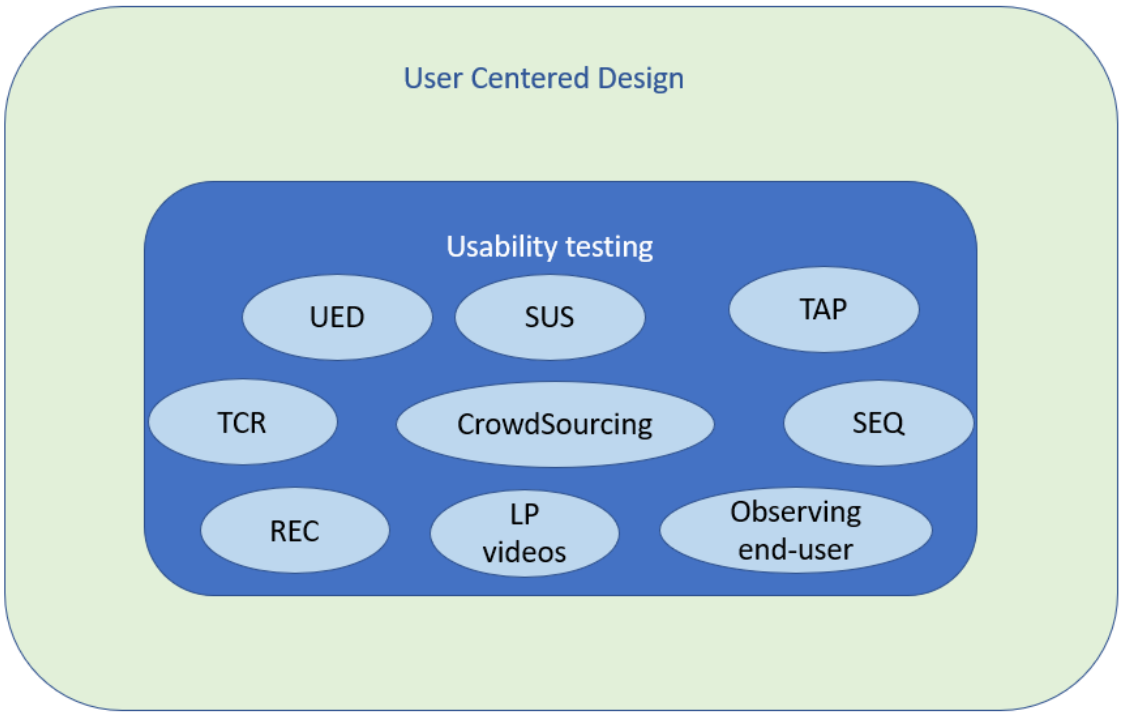

Figure 2. Usability testing is a part of User Centered Design. Sakari Järvelä, 2021.

The contexts may define reasons to use or leave out some methods for the practical part. The understanding of the contexts assists to define a need for the method. It is possible to derive from the researched literature that User-Centered Design (UCD) operates as a top concept for usability testing and there are various usability testing methods that can be utilised (see Fig 2.).

2.1 Contexts and background

The Canadian Pacific Railway (CPR) offers cargo transportation services in North America (Meszaros & Aston, 2006, pp. 1–2). CPR was created to connect the widest inhabited centres in Canada and was founded in 1881. The Web site of the Brown University Library went under restructuring in July 2006 (Bordac & Rainwater, 2008, p. 110). User preferences and needs were the centre of focus in restructuring by the web team to refactor usability. Garousi (2010) had a research on Canadian testing methods

overall. The Canadian province of Alberta was under research of testing methods (Garousi, 2010, pp. 2251–2252). The justification to research in Alberta was that the Canadian software industry has a gap to the needed level of execution. Usability testing has gained more attention in Alberta organisations using experts´ opinions and observations.

According to the research based on 12 different countries with 12 professionals on the usability field, it can be stated that there were forced development and innovation when customers started to use Agile (Nielsen & Madsen, 2012, p. 261). It is stated that the software production and the academic world are having different points of view that are far away from each other, in terms of usability and UX (Ovad & Larsen, 2015, p. 40). According to Ovad and Larsen (2015) the discussion about the different points of view has been ongoing over a decade. Gibson et al. (2016) made a research around usability in an environment of the end users living with dementia. Some usability Assessing methods that define usability level of applications intended for people living with dementia and their caregivers was done in the research (Gibson et al., 2016, p. 1). In the research of Jarraya and Moussa (2018), there was projected a continuous, automatic, and real time usability evaluation for human-system interaction (Jarraya & Moussa, 2018, p. 464). May (2019) explored LP recording as a usability testing method. On June 2018, LP recording search on YouTube gained over 13 million results as a statement of sheer volume popularity of the LP recordings (May, 2019, p. 96).

The software project for CPR was to offer cargo rate information of railcars through web-service-based application when it came to class 1 railroad carriers (Meszaros & Aston, 2006, pp. 1–2). Once the needs were defined, the outcome was to build an application for interline price quotations. The first web site of the Brown University Library was released in 2001 (Bordac & Rainwater, 2008, pp. 111–113). The first version of the web site gained positive feedback from undergraduates and varied feedback from the staff. As time passed, the new added material caused new needs for design and usability causing a frustrating phase as the need of maintenance increased.

Usually, software engineering and the usability testing follow different paths and usability testing does not operate as a part of software engineering (Winter, Rönkkö & Hellman, 2009, p. 108). The main reason leading into this situation could be the multidisciplinary need for the usability testing. The Usability testing practitioners have been qualified more for the usability testing than the software development. In Iceland, system testing was the most traditional testing method when usability testing and performance or load testing were used basically on their level (Larusdottir, Bjarnadottir & Gulliksen, 2010, p. 105). When application is assessed by usability testing, bringing up new features should still be an ongoing process (Tarkkanen et al., 2013, p. 91). In addition, the order in which to accomplish the test tasks and formulate them are important parts of traditional usability testing task formation.

The fully operational application is not mandatory for usability testing since it is not a presentation of the system (Bandi & Heeler, 2013, p. 1). The systematic usability testing approach was used to avoid looking for bugs instead of confirming that the application is effortless to use. Ovad and Larsen (2015) researched collaboration between academia and production. They examined eight companies in 2013 and seven in 2015 in Denmark. Usability testing was done in collaboration with academia in two companies in 2015 (Ovad & Larsen, 2015, pp. 40). Software designers are aware of the importance of usability (Barra et al., 2019, p. 128). Consequently, resources are being used to observe

user behaviour and observing leads to a need to have someone observing. The stated result is that there must be human resources to use in observing. Therefore, user interface usability testing automation was proposed.

There were only a couple of studies about usability integration to Agile (Nielsen & Madsen, 2012, pp. 261–262). Any studies about the Agile's impact on usability production were not found. The traditional status quo was that the usability tests were shortly described at the end of the Agile iteration. The following concepts were made: micro testing, user workshop and a feedback day to describe the one to two days lasting period of usability descriptions. In many cases the usability testing is conducted with a real software that is already developed and then usability issues are fixed after development (Patton, 2008, p. 1). Iterative user testing in action research operates as a base for the researched case study (Bordac & Rainwater, 2008, p. 110). It is supported that one user group may suggest helpful input also in the perspective of other user groups. Novice users may come up with perspectives that could be used as a base to build an easy-to-understand model.

There are problems to deliver quality products meeting the budget and the time criteria (Garousi, 2010, p. 2251–2252, 2257). Therefore, the methods and practices used in the development can be questioned. The focus of the research was on testing the methods overall. There was less time used for usability experiments in 2009 than in 2004. The opinions of experts had more value in 2009 than in 2004. The Agile practises might have had effect on the trend. In Iceland, Software testers did 43% of the usability testing (Larusdottir, Bjarnadottir & Gulliksen, 2010, p. 105). The Lack of time was the main reason for not to proceed with usability testing. The use of usability testing is often low on website development because of the costs and the additional delay it is causing (Liu et al., 2012, p. 1). For this issue, crowdsourcing is explored as an alternative way to engineer usability testing.

The collaboration between academia and companies is possible to see as a proof that collaboration of academia and production can assist to find better UX creation and usability testing approaches (Ovad & Larsen, 2015, p. 48). The study was limited in quantity and that in turn limits the generalization making from the results. The success of a product often depends on understanding the needs and desires of its users and those demands should be taken into consideration in the process called the New Product Design (NPD) (Choi, 2015, pp. 2245–2246). There are several methods with which those demands can be identified, such as gathering opinions or input on a product by revealing some representation of it. The effect of usability testing may appear in various and surprising forms in sophisticated software development (Tarkkanen, Harkke & Reijonen, 2015, p. 301). When evaluating HCI, the usability has been in a centric role with high importance level (Reeves, 2019, p. 1).

The Definition of usability can be stated as an area where user successfully, competently and contently is achieving the goal of the action (Jarraya & Moussa, 2018, pp. 464–465, 467). With those statements in mind, it was possible to use a proof method with a heuristic assessment method as a combined approach for the usability definition. Jarraya and Moussa (2018) examined usability testing in the context of car speed controlling. The Automated Speed Control system was put in a car simulator environment to be monitored in real time. The evaluation in that context was done with 15 end users from various professions with dissimilar information technology skills and diverse ages. An Integration of the Automated Usability Evaluation (AUE) would be useful to achieve in the future

(Hussain et al., 2012, p. 215). It would be beneficial to achieve collaboration between HCI experts and developers in terms of defining the AUE requirements. The collaboration would ensure that both stakeholders are having a common view of the feature requirements.

Domain understanding and IT experience are required from human subjects selected for the usability testing (Bandi & Heeler, 2013, p. 7–8). The risk of lacking domain knowledge was bypassed by giving the required domain knowledge in a form of explanation of the case study. The generalization of these results may not be appropriate as the subjects are not from the same domain than the actual case. An Intuition is not a way to go with the design of online technical communication courses (Hovde, 2015, p. 5). The usability should be evaluated with heuristics and direct measures of the usability. The heuristics and the direct approaches are both important because e-learning system interacts fluently with the end users. The operational accomplishments of the assessors who have found an issue could be a definition for usability findings (Reeves, 2019, pp. 1, 33). The research brought up that usability testing usually expects that there will be findings.

Various tasks for users are being used to evaluate usability (Barra et al., 2019, pp. 132–133). It is also possible to have various actions in each task. Rapid Iterative Testing and Evaluation (RITE) differs from the fixing after development concept by creating paper prototypes and simulating the actions of the end user (Patton, 2008, pp. 1–2). Test-fix-test-fix called cycle reveals the usability issues of the prototypes. When the cycle is repeated until the prototype starts to be usable, the goal is reached. Then the software is easier to be developed in the usable form. The validity and the efficiency of the usability engineering methods have not been in a big role in empirical studies concerning the usability engineering (Liu et al., 2012, p. 2). Online applications or websites can be tested by using the crowdsourcing that offers a great variety of users. The success of a product is determined by its functional needs and few different attributes (Choi, 2015, p. 2244). It can be defined that those attributes can be for instance aesthetics, satisfaction, and acceptability. Those features are especially important when talking about an assistive product in which the functionality is fundamental.

People at the Center of Mobile Application Development (PACMAD) was using specific parts of the usability models to produce an inclusive overall model for mobile application development purposes (Zein, 2016, pp. 335, 339). Effectiveness, efficiency, and mental load conducts the usability itself. When the focus of an application is user-centred, the evaluation of usability is an important aspect (Germanakos, 2018, p. 85). Qualitative and quantitative data were gathered to find the issues and reduce clumsy behaviour. The difference between the usability characteristics evaluation and the requirements on functional operation must be recognized to gain understanding about the whole application behaviour (Eck et al., 2018, p. 91).

In the research by Barra et al. (2019) usability testing begins when an interface designer uploads the interfaces on the PlatoS Server. Then the mockups will be downloaded from PlatoS server with PlatoS Android app onto one's mobile by the interface designer. Furthermore, the interface designer executes the planned actions in proper order to the mockups downloaded before, meanwhile PlatoSapp records action classifications and measures time taken to achieve the goal of the task. The recordings will be uploaded to the PlatoS server. The interfaces are downloaded from the server by end-users who then execute the tasks to create their model of use. There is an automatic comparison of the

usability metrics to spot specific trouble in interface usability so the designer can modify the prototype. Mockup Design activity receives the mockups with problematic performance and the designer has access to Mockup Design activity. Android Studio or programming editor can be used to create interfaces. (Barra et al., 2019, pp. 130–131.)

The logging method is the basis of PlatoS' user testing (Barra et al., 2019, pp. 132–133, 134). The End-user can provide personal data on volunteer base to user group comparison. The goal is that the observers or professionals should not be needed to monitor the usability testing with PlatoS. The usability metrics must be defined to reach that goal. Also, Liu et al. (2012) used user testing logging that was possible to execute remotely. According to them, usability testing can be done remotely via Amazon Mechanical Turk and CrowdFlower tools that are known as crowdsourcing platforms (Liu et al., 2012, p. 1).

The Crowdsourcing platforms offer quick turnarounds, expense savings and a large user base (Liu et al., 2012, p. 1). Liu et al. (2012) organized a study considering crowdsourced usability testing on a graduate school website to see the potential of crowdsourcing. It appears that in the context of technical communication courses, there could be used student feedback, analytic tools, and usability testing (Hovde, 2015, p. 3). When using different tools, for example a quiz that is expected to take ten minutes to execute, can take twenty minutes, but the reason can be unknown. That unknown reason could be found with student feedback, analytic tools, and usability testing.

In the area of users living with dementia, there does not appear a great amount of research (Gibson et al., 2016, p. 2). If the targeted user group has a lower cognitive operation level, traditional techniques of usability testing may not be applicable. For User Experience (UX), the usability testing is seen as a crucial stage of the process (Germanakos, 2018, pp. 85–86). The drawback of usability testing is heavy testing which is usually rather resource consuming (Eck et al., 2018, p. 91). A suggestion to use the product usage data to create usability test scenarios becomes relevant. Requirements engineering can be seen successful when the goals of the functional development are known before the system design is complete. The Mockups of eCommerce mobile application operate as an example to explain the practise of automated usability testing (Barra et al., 2019, pp. 128–130). The Usability testing sessions can be assimilated with LP videos because the content is using TAPs in both contexts (May, 2019, p. 94).

ASP.Net was used to build a browser-based application with user interface (Meszaros & Aston, 2006, pp. 1–2). The document-driven procedures of several inside assistance units and outside the Interline Partners were implemented in eXtreme Programming (XP) methodology to create the application. The Agile method was used for the CPR project. ClearStream Consulting was purchased to offer detailed Agile progress practices and mentoring of the project supervision. In 2006, the new design development program was started (Bordac & Rainwater, 2008, pp. 111–113). A lean design team was required to approach with the planned Agile trial-and-error model. There were organized interviews and observation with the staff who interacted with the users (Bordac & Rainwater, 2008, p. 116). The assumptions on the generalizations about the content and functionality were possible to be made through the researched case studies found. The assumptions were tested after defining who the users are.

There were used mTurk and CrowdFlower to explore the capability of crowdsourcing in usability testing (Liu et al., 2012, pp. 1–3). Tasks can be uploaded through CrowdFlower

on mTurk or some other crowdsourcing channel. The large projects are divided into small tasks in CrowdFlower. The answers of the users were audited through "Gold Units" and if they contained too many mistakes, the answers would be discarded from the results. The only known crowdsourcing company targeting usability testing to the authors, was uTest. The research is a proof of the usability professionals being under demands and high pressure (Nielsen & Madsen, 2012, pp. 261–262). The usability professionals experienced the situation as a threat, but there were benefits to be found when implementing usability testing in Agile. A way of implementing usability testing in Agile was seen most likely to be found by trying different concepts and finding issues and benefits of these concepts.

The mental model shows interactions among the operational units of the application with boxes and lines (Bandi & Heeler, 2013, p. 1–2). An E-R diagram, short of an entity relationship diagram, is used to find data items when designing the user interfaces. The Task description is used to verify that the tasks are supported in simple and complex cases by the user interface. Virtual windows are used on paper or electric view to confirm that the required data is showing. The validity of several components of the input could be examined to improve the decision making during the designing based on the user input (Choi, 2015, pp. 2246, 2248–2249). Choi (2015) made a research about end user input in early phase of product development. The used design artifacts were concept narratives and sketches, appearance models and 3D renderings. To improve the decision making during the designing based on the user input, there is another potential method: using the application of augmented reality. The questionnaire of the Choi's study included four categories: usefulness, ease of use, ease of learning and satisfaction.

In research of Tarkkanen, Harkke and Reijonen (2015) open-ended structure of usability testing offered precise information and a way to find issues that were not identified in the development phase. Especially for sophisticated development background, it is recommended to use open-ended testing methods (Tarkkanen, Harkke & Reijonen, 2015, pp. 301, 311–312). Open-ended approach did not overcome the traditional usability testing, but open-ended approach did elaborate the understanding of the context. To define application usability, there were organised sessions for the end users to gather data, ask questions and observe them acting with the application (Germanakos, 2018, pp. 85–87). The data can be collected with direct questions or by the methods of application that can define the usability dimensions, but both ways are defined by end user's actions. A combination of straight questions and methods with usability dimensions is called a mix of implicit and explicit data.

In research of Barra et al. (2019) the definition of the application usability begins when the user device can run mockups of the Android interfaces. The performance of user model can be compared to the performance of the perfect model that is the one designer wants to achieve. As a result, it is possible to have our focus on differences. PlatoS is an instrument that can find those differences. The interface designer can define the Android mockups with supposed behaviour. The interaction of the mockups is recorded at the user's end. The statistic examination of the log data is automatically analysed to demonstrate usability problems in PlatoS. Usual tools planned for designer to generate interactive mockups do not support user behaviour data gathering. The lack of that data gathering upraises the need of a human observer. (Barra et al., 2019, pp. 128–130.)

The external relationships with alumni and donors, off-campus users with medical school affiliates, general information with novice users, faculty and researchers were the user

profiles defined for the case study (Bordac & Rainwater, 2008, p. 116, 119–121). Generalisations are hard to be made to satisfy all users' preferences. The learning preferences follow the difficulty because people have different assumptions about the use of specified web pages such as library web pages. An easy access to the needed tool is in an essential role when students are using library web pages. If the tasks remain open-ended, the results differ from traditional usability testing tasks (Tarkkanen et al., 2013, pp. 102–103). In the research by Tarkkanen et al. (2013), using open-ended tasks was an effective way to provide results. The traditional usability testing tasks were still required to support the open-ended tasks because open-ended tasks did not fulfil the use of traditional testing.

As a result of the study of Bandi and Heeler (2013), users were not able to finish complex tasks (Bandi & Heeler, 2013, p. 8). The reason is the no un-answered questions. Moreover, the iterations from first to third caused reduction of the design issues. For usability testing method, the results provide further support. More users for more large case study and more complicated tasks concentrating on different areas are planned to extend the whole study. The User experience must be taken care of by online technical communication courses (Hovde, 2015, p. 1). There are perspectives of students and faculty members who design the interface and teach the classes. The effectiveness of learning management systems (LMS) with sufficient user experiences should be the factor that gains focus and prioritizing in the faculty of technical communication.

The observation and the recording of the task accomplishment levels and times in the research of Gibson et al. (2016) were found the most consistent ways to define usability level for dementia context. There were not consistent data coming out from think-aloud methodology. The cognitive load of think-aloud protocol is probably the reason for difficulties in describing the operations. SUS was found unreliable too. In addition, the post-test surveys were found inadmissible with the predominant circumstances. The camera was found more as a misleading than usable observation device for the usability testing. Even the voice records were not usable as the human-computer actions were difficult for the users to describe. (Gibson et al., 2016, pp. 1–2, 6.)

The main conclusion was that traditional usability testing methods are not applicable for people living with dementia (Gibson et al., 2016, p. 6). Conclusion of Germanakos (2018) about usability testing is the outcome that is looked after to find issues in an early phase (Germanakos, 2018, pp. 85–87, 91). The gathered data is usually complicatedly structured, and the Engineering Usability Research Empirical Knowledge and Artifacts (EUREKA) is a paradigm or methodology to analyse empirical data from multiple sessions. Discover, Learn, Act and Monitor are the phases of EUREKA to make easier for researcher to construe the results. The original design of an application usually goes through a lot of changes which causes a loss of the effort made for the application (Barra et al., 2019, pp. 128–130). The differences in the established paths are noticed, and the paths present the tasks visually. There is no need to download the software when using web technologies. The case study of Barra et al. (2019) contains only qualitative data.

One suggestion of usability testing group was to add a guide to instruct the users to find the information they are looking for (Bordac & Rainwater, 2008, pp. 119–121, 124–126). The lack of knowledge in that situation is not seen as the main issue, but it informs that there might be even more challenges in the most basic functionalities for others. The hard-to-find sections were for example library hours, the Brown's collection of electrical resources and the library catalogue. Using web more than once a day and being very

familiar with it, were the statements that five participants came up with in the first usability session. To compare hard cases of more traditional type of usability testing with crowdsourcing, the negative factors of crowdsourcing are the spammers or cheaters who have access to crowdsourcing tools (Liu et al., 2012, pp. 2–3). One approach of spammer is to gain profit from testing without caring about the product or the test results.

The function mini specifications are for producing the full prototype by fulfilling semantic, search and data entry functions for the virtual windows (Bandi & Heeler, 2013, p. 1–2). The virtual windows and data model are made in parallel by using CREDO matrix. The issues appear in the items on the defect list. The inspector is responsible for detecting hypothetical issues in the user interface in heuristic evaluation. The domain knowledge may not be fulfilled by all the inspectors. It is disclosed that the manufacturing of prototypes to gather input from users is expensive and time-consuming (Choi, 2015, pp. 2245–2246). The more a product costs to the company, the more they expect to benefit from the investment. The demonstration of safety and usability requests are challenges for a health gadget (Eck et al., 2018, p. 91). The main case to avoid dangerous conditions is to define the models by the actions of the user. When it comes to the usability characteristics, they can be outside the functional requirements of a product if the requirements are on the functional operation. The result of the research of Barra et al. (2019) suggests a user-centred approach to usability. From The initial phase of development process, the end-users are making effort to the project (Barra et al., 2019, p. 130).

Direct interaction is not happening in crowdsourcing between users and test group (Liu et al., 2012, pp. 3–4). In crowdsourcing, the tasks and the instructions must be carefully designed. To set up crowdsourcing test group, the participants were informed about the examination being a part of an academic study. In the survey, there were also required written answers to avoid random multichoice clicking. The real feedback was rewarded with a bonus to the payment. Choi's article was about the challenges which producers encounter in developing assistive technology products. Whether the before mentioned input is thoroughly comprehended, the designer may concentrate more on the components that are probably going to represent accurately the users' final opinion of a product that is complete (Choi, 2015, p. 2244).

In the research of May (2019), there were selected five videos by using traditional setting of standards with usability heuristics. When assimilating TAP to LP video, there can be found similarities in the explanations for user residents, striving and opinions on the application (May, 2019, pp. 94–95). It is commonly comprehended that to gain better input from a user, the product representation should be made as realistic as possible. The aim in the research of Choi (2015) was to evaluate the statement mentioned above. There is explained that the product representation is conducted by comparing two schemes: the user evaluations of different types of design representations and evaluations of real products which they are based on (Choi, 2015, pp. 2245–2246). It is important to receive feedback on the product when it is at its early stages in concept development.

There have been many occasions where a specific feature or a function have been asked by the users during the development of a product, yet the exact same feature is being criticized in the final product. The products can be mass produced these days and the development stage has been enhanced. However, the success of the products is not necessarily guaranteed. (Choi, 2015, pp. 2245–2246.)

2.2 The Guidelines to execute usability testing

An Accurate test scenario creation requires deep familiarity with the application features, the task presentation, and the used context of the application (Eck et al., 2018, p. 94). The paths of the test tasks must meet the end user's performing when one is using the application. The match of the test tasks must happen in the specific order. In the early phase of prototype usability testing, it is possible to gain early usage data to have the effect on the development (Eck et al., 2018, p. 95).

Without usability testing and the effort on user interface design, in the research of Meszaros and Aston (2006) the end users found a great amount of usability issues in the first release. After all, usability testing is easy to evaluate even if the application itself is not build yet (Meszaros & Aston, 2006, pp. 2–3). Prototyping can be done in a traditional way by using a pen and paper. Bordac and Rainwater (2008) organized the usability testing with the undergraduates and the web design team having usability testing and focus group sessions. The old web site was usability tested and redesign started accordingly in July 2006. The idea was to find what users expected from the web site, what kinds of other sites were used and what way they used the web site (Bordac & Rainwater, 2008, pp. 118, 129–130). As a result, it was possible to state that one user group may offer basic input to gain better usability.

In the research of Hussain, Slany and Holzinger (2009), the quantitative and the qualitative research methods were used with web-based survey to research integration of the usability and user focused methods with the Agile environment (Hussain, Slany & Holzinger, 2009, p. 420). The usability testing was conducted for a website of a graduate school with crowdsourcing (Liu et al., 2012, pp. 3, 5–6). The tasks were assigned for the website that the users had not visited before. The idea was to define the problem areas by the interface and the content of the website. The traditional usability test was completed to compare it to crowdsourcing. For the traditional usability test, there was generated a usability lab setting with volunteer members from the graduate school.

The first research question of Bandi and Heeler (2014) was asked as: If there are no un-replied questions, is it possible for the user to accomplish the key assignments? The second research question was stated as: When using prototypes to fulfil the tasks, what kind of design problems users have? Time is measured in seconds to operate as a metric in research. When operating with given task, another metric is to measure the amount of usability problems (Bandi & Heeler, 2013, p. 3). Using laboratory conditions for mobile device, the usability testing differs from the context of use, and therefore the using context should be taken into consideration. (Wetzlinger et al., 2014, p. 1).

In research of Pillalamarri, Huyett & Abdel-Malek (2015), the system Usability Scale and key performance indicators (KPIs) like click count, ease-of-use measures, step amount and time on task were used to test usability (Pillalamarri, Huyett & Abdel-Malek, 2015, p. 1134). For Agile development, the important view is how the allocated tasks can be distributed flexibly, when usability view is on how the end users will use the application (Alomar et al., 2016, pp. 197–199). The case of the study uses JIRA, AgileZen, VersionOne and ZebraPlan programs as the pragmatic examination target. The most traditional usability testing methods used in the research were observing the end-user, TAP, the video analysis recording (REC), The Single Ease Question (SEQ), the task completion rate (TCR) and SUS (Gibson et al., 2016, pp. 2–3). SEQ uses a 7-point scale to determine the difficulty of the given task. The difficulty is determined Before and after

the execution of the given task. In the research of Alomar et al. (2016), Task Completion was used to measure the usability in the manner of successful usage of the application. The Time on Task was a metric to show how quickly a task can be executed and thus to measure how logical the application is (Alomar et al., 2016, pp. 202–207). The Number of Mouse Clicks measured how easy it is to reach the desired functionality.

In the usability testing of the research of Potts, Nguyen & Turner (2016), there was applied notation system for quick testing. Notations taken from quick set-up requires all preparations of materials ready before starting the test (Potts, Nguyen & Turner, 2016, p. 1). One way to do usability testing is to explain the situation to the end user (Osorio, Aristizábal & Zuluaga, 2016, p. 4). At the beginning, it would be reasonable to explain that the application is under testing, not the ability of the end user. Then the application should be introduced briefly. The practical differentiation between complexity, risk level and functionality diversity can be seen from two examples: A medical imaging system and a smart bottle (Eck et al., 2018, p. 92). The medical imaging system may have great effect on the patient and the smart bottle operates easily with any caregiver to offer right heated liquids.

Four male and four female participants between the age of 25-29 years, were chosen as participants for research of Barra et al. (2019). The participants were required to be smart users who had used eCommerce (Barra et al., 2019, pp. 136–137). The mockups were needed to be analysed with eCommerce. The university of Salerno was a place where the study was performed. Recently, assessing the game usability has been a subject of study to define the heuristics because of the popularity of video games (May, 2019, pp. 96–97). The initial tutoring could be made supported by the complete list of quantifiable standards that varying LP gamers offer. Gibson et al. (2019) usability testing was done by TAP, REC, SUS and observing the end-user as traditional methods that were used as usability testing methods in a context of people with dementia (Gibson et al., 2019, pp. 130–132). The System Usability Scale was used to define the usability from the end-user's point of view (Alomar et al., 2016, pp. 202–207).

The target observations must be on the performed activities, order of using the activities and how long it took to complete it (Eck et al., 2018, p. 96). As a result, from the activity logs, it is possible to come up with models of user behaviour. The created user behaviour models can be used to produce evidence-based usability test scenarios (Eck et al., 2018, p. 97).

In study of Meszaros and Aston (2006), a paper prototype was prepared, and the prepared test tasks were provided for the end user. The application was introduced to the end user in general level because the logic of the product was under research scope (Meszaros & Aston, 2006, pp. 2–4). Acting as a computer behind the application operations, operating as a help system of the application, and an observer of the application usage were the roles played by the software development team members. The feedback from the usability testing was already familiar knowledge for Bordac & Rainwater in their research in 2008. The data collection-based individual usability testing, functional prototypes with search prototypes in focus groups and the revised functional prototype individual usability testing with interviews were the phases completed in the whole testing process (Bordac & Rainwater, 2008, p. 118).

The Online Public Access Catalogue (OPAC) is supposed to search everything from the end user's point of view (Bordac & Rainwater, 2008, pp. 129–131). That causes a

problem in the end user's knowledge about how to access the databases more comprehensively. The usability crew was not able to solve this alone with re-designing. A questionnaire of Hussain, Slany and Holzinger (2009) was based on open and closed questions with multiple choices on a 5-point Likert scale and text field answers. There was also a possibility for an answer that implied that the user does not have the knowledge to answer the question (Hussain, Slany & Holzinger, 2009, p. 420). Overall, 28 questions were asked in the survey. A questionnaire of Hodgetts (2006) was used to find out the degree of satisfaction with the tools used for UED (Hodgetts, 2006, pp. 6–7). A pilot test was used to exercise the test with 11 typical users (Liu et al., 2012, p. 3). The second test was executed with 44 additional test users After the test environment had been modified. There were also changes to the third test that had 50 participants. The used crowdsourcing platform was mTurk. The users were asked to fulfil a survey before the tests and then to complete four tasks on the website.

The traditional usability test and the crowdsourcing usability test were purposefully differentiated from each other to also develop the website to perform more efficiently (Liu et al., 2012, pp. 5–6). In the traditional setting, there was pre informing about the purpose of the tests and a pre-questionnaire. There were five tasks for the participants to complete. Think aloud was used as a procedure to inform the observers how the participants are performing. The process took longer, but it explained issues in more comprehensive way. At the end of the traditional usability test, there were six qualitative questions asked for the participants. The questions were set in a way that the participants were able to reflect their own experiences when answering them. The timespan of the traditional usability test was from 20 to 40 minutes. Another usability testing setup was The "Bubba's Healthy Snack Project" that was chosen as a case study to fit in the scope of the research by Bandi and Heeler (2013). The norms to choose the participants for the usability testing were around the idea of the meant end users (Bandi & Heeler, 2013, pp. 3–5). Because of that limitation, the end users had to have Information Technology background and domain knowledge of the given case study. The snack project was explained to ten selected graduate Information Technology students to gain domain knowledge of the snack project. After creating the required deliverables, the usability testing was organized through a systematic procedure. the user, the facilitator, the log keeper, and the computer were the roles played during the usability testing.

To set the concept before the evaluation of tools JIRA, AgileZen, VersionOne and ZebraPlan, it is reasonable to set the evaluation criteria to evaluate the tools (Alomar et al., 2016, pp. 197–199). Few of the criteria are the general user-responsiveness, the applicability of the presented symbols and the readability of the written texts of the tool under evaluation. The reached benefit after the study is to offer describing data for the developers of the tools that were examined in the usability research. The developers will be able to see the usable and the unusable parts of the applications and thus do modifications to have a more usable application. All the usability parameters should not be judged equally important as the usability depends on the end-user's needs and the domain requirements of the target systems. The comments the observer makes, should be made from the test user's point of view (Potts, Nguyen & Turner, 2016, pp. 1–2). A numbering system was required to keep track on which tasks the test users are working on. The issues should be marked with different colouring in the notes. To have reliable background for notes, one must use quotes from the test users.

According to Osorio, Aristizábal and Zuluaga (2016) the test organizer should guide the end user to think out loud. The questions should be asked before the end user starts to test

the application then the following act is to observe. The fifth part is to answer the questions of the end user and inform about the observations (Osorio, Aristizábal & Zuluaga, 2016, p. 4). The experience of the end user should be asked before the final phase of utilizing the results. In the research of Barra et al. (2019) the developer used the interface mockups to create comparable model before the tests begun. There were three iterations for the evaluation (Barra et al., 2019, pp. 136–137, 141). The assessment was created by a statistical method with an automatic discovery of differences between the end-users' and the developer's behaviour. The outcomes were reassuring based on the example test process. May (2019) then put the test organizer in a role where showing ordinary characteristics were chosen to establish a rule to assess the usability of a game while watching LP videos at the core of the research when reviewing the heuristic guidelines. The six characteristics of the game's usability are: 1. The availability of status 2. The adherence to conventions 3. The customizability 4. The accessibility of help 5. The comprehensibility 6. The assistance with user recall (May, 2019, pp. 96–97).

The study of Bandi and Heeler (2013) was conducted using three iterations. In the first one, only one user took part in the usability testing (Bandi & Heeler, 2013, pp. 3–5). Likewise, in the study of Bandi and Heeler (2013), Liu et al. (2012) conducted usability research where three iterations were used. In the research of Liu et al. (2012), the pilot test had been planned to take 10 minutes and all the results were conducted in less than three hours (Liu et al., 2012, p. 4). The total cost of the pilot test was 2,93 USD. The realisation of time on survey was around 13 minutes in average. The survey was completed quicker with a background of a Bachelor's degree or higher education level compared to an associate degree or lower. The textual answers were short, and the specifics were not included in the responses. The pilot test did not contain spammers.

For Bandi and Heeler (2013), the second round contained two users, and the last one seven users (Bandi & Heeler, 2013, pp. 3–5). For Liu et al. (2012), the second test was put together by giving four tasks for the participants. It was asked from the test users to provide the information about time on task by clocking how long it takes to complete the task (Liu et al., 2012, pp. 4–5). When the tasks had been completed, the asked questions required answers in the textual format. The planned time span for the survey with tasks was between 15 and 20 minutes. The total cost was 23,41 USD with 44 participants. There were 14 participants suspected as spammers as their responses were random. In the task 3, the participants were instructed to visit a website to evaluate its usability, but the spammers did not visit the website. In usability testing, the challenge is that any fault can be considered as a usability issue. The visit to the offered website could be verified by asking the first word from the website, but it would not solve the issue of participants not paying attention to the website. The time spent on task was not matching the time taken for the whole survey in one case. The questions requiring textual format answers performed better because the questions were more detailed. When compared to the pilot test, the value of the answers was higher.

In the study of Bandi and Heeler (2013), the first and the second iterations were conducted to create a new edition of the prototype for the third iteration (Bandi & Heeler, 2013, pp. 3–5). Liu et al. (2012) had an approach where the basics of the third round were equal to the second round. The time on task was replaced with scaled questions from one to five, to state the difficulty level of the task (Liu et al., 2012, p. 5). There was also an optional field to comment the task itself. At the end of the survey there was used SUS. The planned timespan for a participant to finish the survey was between 20 and 25 minutes. The overall cost of the third round was 55 USD. It took less than an hour to collect the results. Only

4 of the participants were considered as spammers. Also, Alomar et al. (2016) used SUS in their study. For conducting a task-based usability evaluation, the Morea usability testing application was applied by Alomar et al. (2016). The software development administration tools were the measurement targets the evaluation checklist was made for. (Alomar et al., 2016, p. 199) The experience with the selected tools and the participants' knowledge of the software industry were defined by using a pre-session questionnaire. The SUS questionnaire was used to gain understanding of the general usability of the tools researched in the study. To measure the usability of the tools from the project management's point of view, there was used a 5-point Likert scale questionnaire.

Bandi and Heeler (2015) thought that the third iteration contained less problems than the first two iterations. The research outcome was that there should be limitations between the organizers and the users in the testing process. The hard and easy tasks should be separated to the user so the user will spend a proper amount of time for each task. The usability testing reduced costs in the process of developing the software. The maintenance of the software becomes easier because usability testing makes the application to be easier to understand and re-use. The life cycle of software development should be put together with usability testing whenever it is possible. The users that are not suitable enough for testing, should not be selected because it may have effect on the results. Iterations are recommendable for usability testing because corrections can be done to the prototype between iterations. (Bandi & Heeler, 2013, p. 6–7.) To compare the amount of iterations, Pillalamarri, Huyett and Abdel-Malek (2015) came up with a study where eight iterations were found suitable amount to conduct usability testing (Pillalamarri, Huyett & Abdel-Malek, 2015, p. 1136).

To gain domain knowledge for the users, the case study behind the research was explained to users. The inspection data collection was used to explore how the users perform with given assignments. To gain deeper understanding of the phenomenon under study, the inspection is the way to proceed. The tasks were introduced one after another for the user. The suggestions and opinions were listened after the task had been done. Think-aloud was a requested protocol for the user to follow as the session was recorded on a video for the further exploring. A Likert scale from one to five was used to offer quantitative data about how the prototype performed with giving access to the required data. (Bandi & Heeler, 2013, pp. 3–5.)

Different tabular forms were used to represent the gathered data (Bandi & Heeler, 2013, p. 6–7). In the case of Research Question 1 with no un-answered questions, there were not participants who completed all the tasks. The Participants were not in need of asking questions to accomplish easy tasks and most of their questions were about complex tasks in which there appeared to be design issues. In the complex task cases with issues, the Likert scale opinions of users were less than 3 when in other cases it was 5. In the case of Research Question 2, looking for problems came out with several issues. For the usability testing, the participants expected output with rich functionality that is not required for usability testing. In study of Alomar et al. (2016), Nielsen's heuristics were used to examine the usability of the Agile project management tools' interfaces to scale usability. Few different heuristics were defined to evaluate from the project management's point of view for management-specific tasks (Alomar et al., 2016, p. 200). The key tasks were executed to grade tools, and specified heuristics were used to evaluate the interfaces.

There were 17 tasks to perform in JIRA, VersionOne, AgileZen and ZebraPlan programs for eight participants in the research Pre-session Questionnaire, the Task Completion, the

Time on Task, the Number of Mouse Clicks, the System Usability Scale, the Second Post-session Questionnaire, the Observed Usability Problems and the Heuristics Evaluation were the executed procedures to evaluate the usability of the target management tools (Alomar et al., 2016, pp. 202–207). The Pre-session Questionnaire was executed to gather information about the participants' familiarity with the software development. In the second Post-Session Observed Usability Problems were the problems that the developers identified from the interaction between the end-user and the application. The Heuristics Evaluation rated the tools against the direct feedback, The Real-world mapping, the consistency with the standardisation, the easy to use with the learnability, the layout with the organisation, the flexibility, and streamlining the experience with the visibility and the clarity. In the research of Hodgetts (2006), the sophistication required for every type of creative task and the sophistication needed in the lifecycle of the product were found handy when UED group came up with that kind of suggestions (Hodgetts, 2006, pp. 6–7). When the UED specialists conducted their activities to support the team, the specialists felt as parts of the project community which was found important.

A Random test scenario creation does not offer the warranty to assume that all the activities of development improvements are being covered (Eck et al., 2018, p. 98). This may cause some unproductive usability testing. To avoid the unproductivity, it is recommended to use the evidence-based user behaviour models. The testing can be approached by using a set of mandatory tasks or a tool-supported interactive guidance by a usability engineer on a test scenario creation. According to Hodgetts (2006), some of the UED specialists are capable in only one kind of action (Hodgetts, 2006, p. 4). The culture of the organization might have an influence on the capability in different kinds of actions when many specialists resist different type of activities than the ones identified for them.

After the usability testing was executed with the test end-users, the test end-users were brought together in a group meeting to elaborate their experiences with each other (Meszaros & Aston, 2006, pp. 2–4). The Results of the comments in the meeting showed that there appeared new user stories and new kinds of training plans among the users. For software development, a need was found to add many features back that were earlier considered out of the scope. The duplicate data entrance was a regular unwanted feature. Some of the lower priority features were left out from the project to add features that improved implementation between the old and the new application. If the software development team had utilized only the theory, the end-user's needs would have been missed. Lines and post-it notes were used to link the screens and to show the operating flow. It made also possible to easily add and prioritize the missing stories. The well-defined roles matched and operated with each other in the project (Bordac & Rainwater, 2008, pp. 129–131). Operating well together with dedication was a key to success for the usability testing team. An application does not require all the content to be usability tested. Hiring people outside the library as a limited-duration employee was useful for the project. A never-ending process is a word to describe the redesigning of the library web page.

Usability testing can also be done by manufacturing it with automation engineering (Wetzlinger et al., 2014, p. 5). When that context is integrated with the user interaction observing and the usability analyzation, more research is required. In the study of Alomar et al. (2016) ten out of 17 tasks were successfully completed by all the participants in the best case of the usability when using ZebraPlan (Alomar et al., 2016, pp. 203–204). In

the table of the article, eight out of 17 tasks were successfully completed by all the participants.

Barra et al. (2019) found that the result of their research suggests a user-centred approach to usability. From the initial phase of development process, the end-users are making effort to the project (Barra et al., 2019, pp. 130–131). An Interface designer uploads the interfaces on the PlatoS Server. Then the mockups will be downloaded from the PlatoS server with the PlatoS Android app by the interface designer onto one's mobile. Furthermore, the interface designer accomplishes the planned actions in proper order to the mockups downloaded before, meanwhile PlatoSapp records the action classifications and measures the time taken to achieve the goal of the task. The recordings will be uploaded to the PlatoS server. The interfaces are downloaded from the server by end-users and then they execute the tasks to create their model of use. There is an automatic comparison of the usability metrics to spot specific trouble in the interface usability so the designer can modify the prototype. The Mockup Design activity receives the mockups with problematic performance and the designer has access to the Mockup Design activity. The results of the usability measuring came up with results that SUS, REC and TAP are not applicable measuring tools in the context of an end user group living with dementia (Gibson et al., 2019, pp. 141–142).

2.3 The Benefits and the drawbacks of the usability testing methods

The Crowdsourcing platforms offer an easy way to access the usability testing labour force. In the way of traditional usability testing, the process might take days or weeks from beginning to end. Setting up the processes, the time taken for the greetings and the traveling expenses can be reduced when crowdsourcing. Crowdsourcing also reduces the costs in a project. The number of iterations can be higher in crowdsourcing because it can save costs. On the other hand, the traditional lab usability testing participant offers more quantity in the feedback. The quality of the crowdsourcing feedback can be seen lower than in the traditional usability testing. The communication between the organizer and the testing participants is not offered in mTurk, and there is no way to use think aloud protocol. In addition, mTurk has spammers in their participant pool. (Liu et al., 2012, p. 6–7.)

In the study of Drew et al. (2018) the following usability testing disadvantages are identified: the subjective user perception, misunderstanding of the items by the researchers or the users of the SUS, missing of the improvements for the usability of some features of the application and the validation of the correctness so that the user experience criticises the correct feature (Drew et al., 2018, p. 358). On the other hand, Rajanen and Iivari (2007) state in their study that when focusing on usability as a tool, sales will rise with better result on marketing (Rajanen & Iivari, 2007, p. 521). Reason behind the growing sales is the improved convincing of the customer. Secondly, enterprise gains benefit from better customer gratification. Usability as a focus will rise development costs, but those costs will appear over the time to gain the usability.

Crowdsourcing can offer important data about usability issues even though the quality of the results is lower than in laboratory conditions (Liu et al., 2012, p. 9). When usability test team is limited by time and money situation, the crowdsourcing is a way to organize

tests. To have less spam and to get useful results, the designing must be done cautiously. Choi (2015) states that different design artifacts tend to have the different level of accuracy and therefore there should be variance in the evaluation compared to the actual equivalent product (Choi, 2015, p. 2250). These differences in evaluation prove that this method is useful in examining the users' understanding of a product based on a design artifact. The use of digital interaction is effective in the evaluation by allowing several iterations of digital interfaces which can then be further tested and developed.

2.4 Improvements for usability testing process

The outcomes from few experiences with UED groups do outline the statements in general form (Hodgetts, 2006, p. 3). In research of Hodgetts (2006), the UED specialists were set as members of the Agile project teams, but still the UED work was conducted in UED organization to confirm by traditional project management practises that the work was coordinated properly. Usability testing is needed in wide-ranging areas of applications (Larusdottir, Bjarnadottir & Gulliksen, 2010, p. 108). However, the Scrum sprints are short, and they provide small progress to the project in one sprint. The small progress leads into a situation where usability testing is complicated to take in use.

An invention is useful when it results to the fact that expected task can be anticipated without obstacles or delay (Eck et al., 2018, pp. 88 –90). The User-Centred Design (UCD) looks at the development of the product from the user's perspective, and step by step unites the usability to match with the logic of the user. A Specific case of Internet Start-Up organizations had unified UED specialists in the development society (Hodgetts, 2006, p. 3). The best-case scenario is that the team members offer the needed, specific knowledge to the project. but sometimes there is a need for adjustments in the team to proceed.

The setting of the tests by people around the project is causing an issue of insufficient test coverage (Eck et al., 2018, pp. 88 –90). The problem is usually on setting very few scenarios without notifying the other functionalities that may have a negative effect on the application behaviour. One way to avoid a lack of the test coverage from the test users is to find representative test users to define the usability. The ideal case is to find out which factor in the representative group has the greatest effect on its test results. Then it is recommended to get a wide range of representatives in that group with this specific factor. To define usability, Drew et al. (2018) brings up that there exists a questionnaire to describe the System Usability Scale (SUS) (Drew et al., 2018, p. 357). The questions in the questionnaire are valued in the contract stage related to the declarations that include a multiplicity of usability characteristics. The needed usability characteristics contain estimations about the system's complexity and the amount of support or training required to reach effective usage coverage of the system.

The Data-driven approach is a suggested way to be taken into use to create a creation of the usability test scenarios (Eck et al., 2018, pp. 88 –90). The main phases are: collecting the data, creating the user activity log from the data, forming the user behavior model with process mining, and forming the test scenario based on the models. When it comes to the usability evaluation, there are various advantages reached through SUS (Drew et al., 2018, p. 357). SUS is found easy to respond from the participant's point of view and additionally, the questionnaire is seen quick, free, and managing friendly to execute.

For qualitative usability evaluation May (2019) states that when the natural context has been replicated, and the tester is guided to use TAPs, the ideal test condition is being created. Next, introducing TAPs to the people who are unaware of the topic, is being observed. The LP videos show that those people often cooperate with the system and by doing so, they show how the actions of the software can be observed. The recordings from the testing the system allow avoiding the interaction with the tester during the testing. It should be paid attention on that there is a possibility of manipulation and distortion when watching LP videos. (May, 2019, pp. 95–96.)

2.5 Bringing usability testing in the Agile and the Scrum environments

In the Agile software development environment, there are tools supporting the automated and the distributed usability testing with mockup kind of prototypes (Hosseini-Khayat, Hellmann & Maurer, 2003, pp. 59–60). If usability lacks in quality, it has negative effect on the usage of the application. The integration of a diverse variety of non-coding subjects in main Agile teams has been a great challenge for the Agile process (Hodgetts, 2006, p. 1). In the beginning, the subjects that were considered, were testing and architecture, which were followed by the database development and the user experience design (UED). The UED specialists were Pressured into changing their old habits to the new concepts and procedures that Agile came up with on the evolution of the organisations.

Taking into use the usability and the user focused design methods in Agile software development environments, have been left behind in the evolution of software development (Hussain, Slany & Holzinger, 2009, p. 416–417). Integrating usability and user focused methods in Agile environments have been researched widely. There is a possibility to create a hybrid method using Agile and UCD to benefit from both methods (Losada, 2018, pp. 1–2). The existence of differing philosophies of the AUCDI has pointed out challenges in different approaches. Adding the usability testing in Agile environment has been an important milestone for the development processes (Hosseini-Khayat, Hellmann & Maurer, 2003, pp. 59–60). The mockup kind of prototyping could reduce costs and resources required for the development process. WOz is a mockup kind of prototype to show the end user how one will act with the final application itself.

The main challenges are coordinating the efforts of UCD personnel, the lack of documents, the desertion of non-functional requirements, the lack of time, the difficult prioritising process, and the modularization. Agile allows a small package delivery in short periods. Thus, the delivery schedule is maintained. Subsequently, UCD operates before the beginning of the development to study and evaluate the needed effort. The Agile-UCD projects should not be considered as equals in terms of efforts required. The use cases and user stories are the most common suggestions to be used when integrating UCD to Agile. From the study, there aroused a suggestion to use the User Objectivities as an artifact. The User Objectivities can define the development and the typology for each project. (Losada, 2018, pp. 1–2.)

The structure of Agile development environment is easy to get close with the usability or the human-computer interaction (HCI) based methods. The results of the survey in study of Hussain, Slany and Holzinger (2009) supported that the development process gained more value when the integration was done (Hussain, Slany & Holzinger, 2009, pp. 423–424). The satisfaction of the end-user had aroused with better usability and quality

reached by the integration. The Remote Synchronous User Testing (RS) is a way to integrate usability testing in the Scrum process (Lizano, Sandoval & Stage, 2014, p. 500, 503). It is possible to have smaller usability evaluations operating as an iterative process is currently going on with Scrum. The evaluation can be done by using only two users at a time.

There is not an agreed Agile-UCD hybrid method that defines what parts of UCD should be considered when the integration is completed (Losada, 2018, p. 7). The literature supports that there are several ways to do the integration between UCD and Agile. When project allows to define requirements in the beginning, the integration of UCD can be done in up-front style. If the project does not allow such an act, the requirements must be collected and updated during the project. The ActiveStory Enhanced (ASE) is an open source selection of instruments (Hosseini-Khayat, Hellmann & Maurer, 2003, pp. 62–65). The basic idea is to make easier for Agile squads to build prototypes. ASE gives opportunity to gather usability data remotely. With the Prototype Designer of ASE, developer reaches freedom and flexibility to the design. GUI building is a weak point of ASE.

After every Agile iteration, it is possible to organize a usability testing. The researchers' observations and the end-user's feedback can be used to evaluate an application (Tortor et al., 2019, pp. 66). Suitability, the proper operating, the easiness of using and the easiness of learning are the concepts that usability testing is made to define. It was concluded that when integrating usability testing with Scrum, there is shorter reaction time on the software quality (Rahayu et al., 2016, p. 5). When the interval of feedback is shorter, the correction of the software is completed faster.

3. The Research Method

Here is explained and justified the chosen main categories used for the study. In the beginning, the usage of the qualitative research method is justified. The case study and the action research methods are defined shortly. Finally, the utilizations are being explained based on the literature part. It is also taken into consideration how the triangulation was used to conduct the research itself.

3.1 Reasoning for the Chosen Approach

Finding out what the individuals state and how they act, operates as a purpose of qualitative research (Myers, 2013, p. 5). The main aspect of understanding the reasons behind human actions is to understand the research environment. In most of the cases, the conversations with the individuals are a key to gain comprehensive understanding. It is argued that the reason behind human actions is impractical to explain without understanding the context.

When looking at the purpose for what the qualitative and the quantitative research methods were created, there was made distinctions (Myers, 2013, pp. 7–9). For natural sciences, the natural phenomena were the original field to be studied through the quantitative research methods. For social sciences, the cultural phenomena were the field for the qualitative research methods. There are use and need for both, the qualitative and the quantitative research methodology. The drawbacks and the benefits can be found from each approach. If one is in a need to examine a topic in its complexity, it is the main reason to choose a qualitative research method. The difficulty to research various residents is the weak spot of the qualitative research. On the other hand, the social and the cultural aspects are abandoned or handled in a shallow way in the quantitative research.

In research, the term triangulation can be described as a combination of different research methods (Myers, 2013, pp. 9–10). The triangulation offers a better overall understanding of the topic. With triangulation, the data can be triangulated from different research methods and data gatherings. Using two or more research methods in one research is not very common. The combination of research methods is found difficult to proceed with.

3.2 The Tenets of Case Study Research

Teaching and research are the two greatest areas where case studies are being widely used (Myers, 2013, p. 73–74). The adequacy of a theory or a theory proposition is assessed by the research cases which, in turn, work as empirical evidence for the topic. For a specific area, there is accomplished contribution by the research cases. The research case requires detailed knowledge about the topic for the author. In a study, there must be referring and knowledge with the latest research around the topic. Other scientists must be convinced that the contribution is new around the topic the case study is written about, and the report must be believable. The report needs to be interesting for the other scientists too.

Factors, problems, or relevant characteristics can be found by the research case study to generalize the results for the related types of situations (Myers, 2013, p. 75–76). Ways to

use the research case study with current literature around the topic are to develop underlying the rationalizations, evaluating the hypotheses or testing the hypotheses. Meanwhile, the first type of purpose for the case study is to build a new hypothesis. The central aspect of the case study is: What was the topic under research? The aspect of what was researched is followed by not defining in which stage of topic the research can be executed. *Why* and *how* are often asked questions in a research case study.

The main benefit of the properly written case study research is that it gets recognized by most of the researchers (Myers, 2013, pp. 82–83). Secondly comes the benefit related to the complicated real-life conditions that can be researched by case study. There are four burdens in research case study. Firstly, it is hard to gain access to a specific company having the specific need. Secondly, the researcher does not have control over the situation. Thirdly, the focus of the researcher may be misled in meaningless issues. Finally, a research case study is a time-consuming process from beginning to end. The sum of burdens conducts the conclusion that doing research case study is for the ones with necessary people abilities, dedication, and passion around the topic.

3.3 The Fundamentals of Action Research

Scientific knowledge can be grown by explaining practical challenges, which is the target of the action research (Myers, 2013, pp. 59–61). In the action research, one is looking forward to creating an organizational change while studying the process. Creating a change during the process differs from other research methods. The researchers and the subjects are in a deep collaboration during the action research. Iterating is a usual way to operate in the action research.

As a positive impact, the pragmatic significance of business research can be confirmed by the action research (Myers, 2013, pp. 65–66). A burden using the action research is caused by the fact that many researchers find it demanding to manage the simultaneously ongoing research and action. Another weakness is a trend for researchers to exaggerate the significance of the interference in the research. The risk is the third weakness of the action research as the delays in real world projects may cause failure for the action research. The failure of an action research is not disadvantageous because the failure may teach more than the success. The viability of an action research may suffer from delays.

There needs to be a problem that requires solving to fulfil the part of an action (Myers, 2013, p. 66). The problem must be important to be solved for the company. From the company's point of view, it is also needed to be open for tricky questions from the researcher. The Researcher in the specific field of science must have a need for the practical problem to be solved, and then it is possible to fulfil the part of research.

3.4 Triangulation

For this research, the case study was the main approach in the mean of gaining comprehensive expertise in the theme of usability testing. The comprehensive expertise was gathered through the literature review. The use of that expertise was achieved by selecting approaches used in the scientific literature. The selected approaches were confirmed to fit in the practical project before taken into use.

UCD operates as the starting point of the qualitative data collection. In the first sprint UCD is a package that will be taken in use with few tools of the package. To match the application's functionalities with the minds of the users, it is reasonable to find representative test users to have a decent test coverage (Eck et al., 2018, pp. 88–90). The first approach to use is UED for gaining opinions on the application through an interview. The second data collection pinpoint was TAP. TAP required a natural context and a test user guided to use TAPs (May, 2019, pp. 95–96). The third pinpoint was SUS that offered a quantitative type of data. The multiplicity of usability characteristics can be defined through SUS (Drew et al., 2018, p. 357). The qualitative questions were queried in the interview phase that was performed in the first iteration. The TAP phase was applied during the second iteration. A Small amount of quantitative data was collected in the SUS phase. The research construction in several phases was justified by the need of getting response to the research questions. In this thesis, trying and finding the issues and the benefits is central, as Nielsen and Madsen (2012) mentioned to be the way to research usability testing.

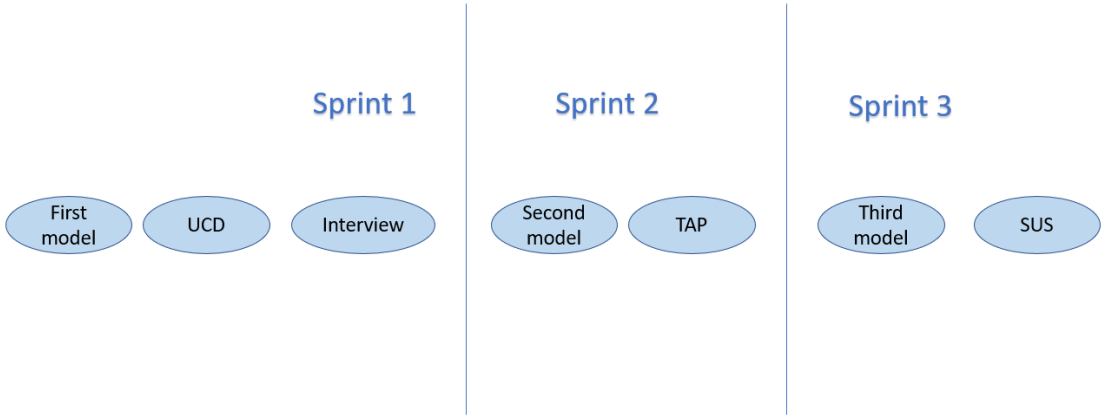

Figure 3. Usability testing methods taken into use with Agile. Sakari Järvelä, 2020.

The triangulation with the action research was achieved by taking the iterations of the action research in the Agile sprint iterations. An Action research also requires changes for the process itself. When the formerly discussed action research and Agile are combined, the sprint can operate as an iteration, and interview, TAP or SUS as a change in the process.

The Change from interview to TAP, and the change from TAP to SUS should have a perspective changing effect on the process (see Fig 3.). The changes were proceeding at the same time the sprint changed. In conclusion, the used research method was a case study research with triangulation to an action research which utilizes the research tools found from the literature review.

4. Implementation

This chapter explains how the research was carried out. First, the background about the company and the usability testing is explained shortly. The created research method of chapter 3 was used to conduct the research which is explained sprint by sprint after background.

The goal of the chapter was to explain how this research can be reproduced. Semi structured interview questions were declared, TAP tasks, and SUS tasks were defined. SUS scale was shown to create understanding about how the system's usability was defined at the end of the project.

4.1 About the Background

Sovelluskehittäjät Ryhmä is a software development company that offers applications for logistics, artificial intelligence, and internet of things areas. One of the services is software development itself. Usability testing was an area with need to widen software development expertise.

In Sovelluskehittäjät ryhmä there are lots of applications created with graphical user interfaces under development. In such environment, the role of the usability assessment is to operate by improving the user experience. The comprehensive software development production was a match with the usability testing research.

The knowledge about the usability testing existence was in the company already. It can be said that usability testing was an option not taken in use. The research about different usability testing methods was a way to see what usability testing can offer as an operative basis for software development.

4.2 Conducting the Questions for the First Sprint

There was conducted a consent form about data gathering to get a permission from test users for data gathering (See Appendix A.). The test user candidates were approached by email. Email informed that usability testing is data gathering for the purposes of this thesis and similar type of research. Candidates were informed that they are allowed to ask questions about the research and they may stop participating any moment they want to.

This research begun with collecting data about the first model of the application under test. The data was collected through an interview, as the research method was explained earlier in this document. Collection was progressed in a semi structured interview mode, that means it was possible to ask questions outside the defined ones during interviews. Semi structured interview has an advantage in that the interview may appear to be more like a conversation than being forced in question answer type structure.

In this research, there were used eight questions that were defined before the interview. Table 1 shows the used structured questions. The questions were set in as generic form as possible because in the starting situation researcher was purposefully left out of knowledge about the application itself. Leaving researcher out of the general knowledge

about researched topic eliminated several questions that could be guided to be asked with the domain knowledge. The guided type of the questions was not totally left out because of the semi structured interview type, but the appearance was forced to simultaneously with the ongoing research.

Table 1. The defined questions for the semi structured interview.

You are about to use the application under test. How do you begin?
Which features do you use regularly?
How do you use these features?
How would you describe the usability of the first model?
Are there some features you experienced difficult to be found?
Are there some features you experienced difficult to use?
What would you not change in the application?
Do you have any questions about the application or the research itself?

The questions were set in a form that should define the application behaviour itself. This way the research can define qualitative description about the environment itself. Overall, the questions should fit in more than just the case of this research.

When the question set was formed, the interview was organized in a way that the test user had possibility to show different use cases of the application. Organizing an interview during this research was another milestone to accomplish. Finnish Government had recommended to organize events remotely if possible. Therefore, actual real-life interviews were not the highest priority. This was solved by using remote meetings.

4.3 Setting Up Tasks for the Second Sprint

The results of the first sprint caused a change towards the planned research method structure. The application model was found complex enough to cover more time span compared to the planned sprint schedules. Therefore, there was found a need to split the model into smaller pieces. Change was that instead of generating entire usability model to be assessed between sprints there was two sections of the usability model to be assessed. One section of the usability model was assessed with usability testing using TAP in second sprint and another section of the usability model was assessed in the third sprint with SUS.

Test users were informed that the application is under the test in the beginning of the second phase when applying TAP usability testing method. Moreover, test users were informed about that any confusion the usability model may cause is not a user error. TAP was conducted through TEAMS where the test user shared one’s screen for development

team to observe. The TAP was explained for the users before beginning the actual test scenario. The explanation contained information about that there will be five tasks for the test user to solve (See Table 2.). It was also explained that the purpose is to describe the execution of the task in detailed manner when proceeding with the task. Example questions to answer during the explanation were: What was observed, what is the feeling about the material available for the moment, what is decided to be done after the observations, what seems to be happening when decisions are done etc. Test users were informed that the ideal case is to have story telling kind of speak before, during and after the actions. Development team shared a link to the current usability model that was conducted with Adobe XD.

Table 2. TAP tasks.

1.	Show where it is possible to find all the 11 targets of the environment.
2.	Find test target named Target 6.
3.	Show tree view and targets of test layer Target location 3.
4.	Activate new target with a name Demo target in the test layer 3.
5.	Inspect information of the Demo target.

4.4 Conducting Tasks for the Third Sprint

The research method structure change was affecting on the scope of the third sprint. The edit list view was the assessed part of the application in the third sprint. The third sprint was focusing on SUS material based on the tasks test users proceeded with (See Table 3.).

Table 3. SUS tasks.

1.	Move in Test Targets list view.
2.	Move in Demo Targets layer.
3.	Find Demo File.
4.	Move to edit configurations of the Demo File.

The first test user was informed that the focus will be on SUS pattern and the tasks are created for test user to get known with the new usability model. This was found as a wrong choice because TAP's descriptive proceeding with tasks gave more test user point of view material about that how the application appears to be. One more change during the research process was done and rest of the test users were informed to follow the known TAP procedure when executing tasks. The TAP procedure was explained in same form than in the second sprint. Used SUS table can be found from Appendix B.

5. Findings

This chapter summarizes the results of the interview, TAP phase and SUS phase. The usability models are shown, and application usage is explained. Results and required usability model changes are described to follow up the success of the process.

5.1 Interview in the First Sprint

The usability model contained functionalities for material handling. The user interface's starting page was set for the simple action selection manner. The first automatic selection was Target materials page (see Fig 4.). In the first model, functionalities edit list (four pointed star), edit top layer (triangle) and targets (five pointed star) were plain icons with the tooltip functionality. Material was controlled from the list view or edited by the top layer. Top layer editing made it possible to inherit material in all the layers when list editing offered more precise controlling on what is the exact target or lower target group for the chosen material.

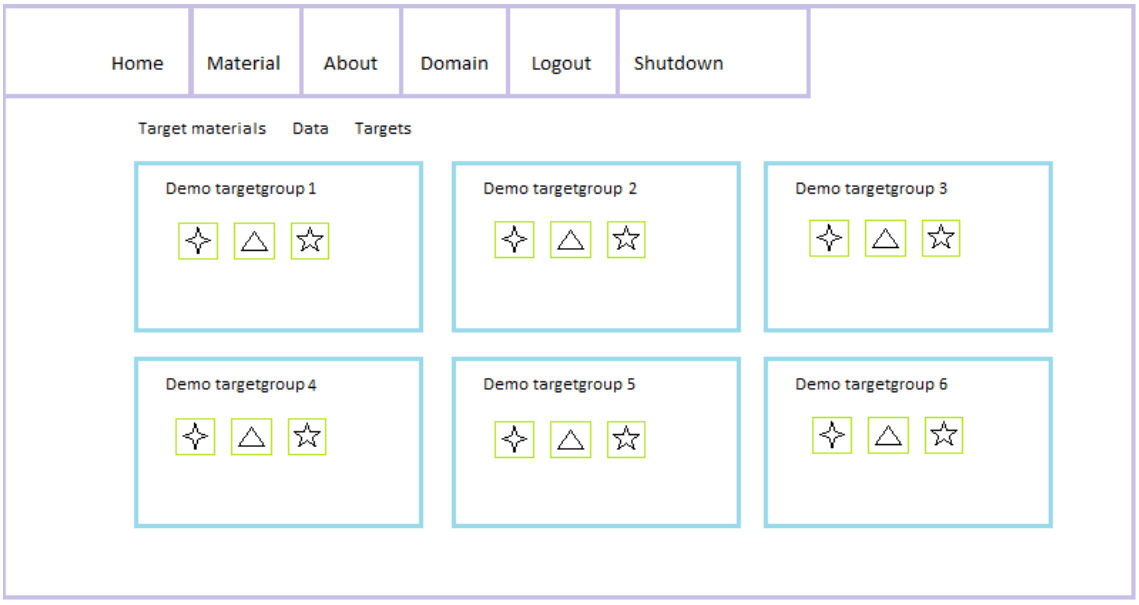

Figure 4. Home view of the first usability model.

The interview results explained that test users usually begin to use the application by logging in and then selecting the top layer editing. In edit top layer test user adds material that is time framed for specific period in all the levels of the targets. Results informed that the time framing is a good feature. The edit list functionality was also used, but rarely. Targets were inspected regularly. Targets might be operative or inoperative, but it was possible to set the targets operative if there was a connection available. Tooltip availability for the icon type of the selections was a possible improvement for the new users. Test users explained that the top layer editing brought ease of use feeling with the software. One complicated process was to add material from edit list menu, because there were many phases, and adding material was experienced as a slow process. Test user would not change the top layer material addition or the targets availability state monitoring existence. One observation about the application usage was that when time

frame was selected, test user was required to click away from the date select to verify the date selection. Double click might make the verification process smoother.

5.1.1 Top layer editing

The first model opened the top layer report view from the home view by clicking “Edit top layer” button. Top layer report view showed current material appearing time periods and had + icon to add new timeframe (see Fig 5.). Title was describing the main idea of the created timeframe when description contained more detailed explanation about the material.

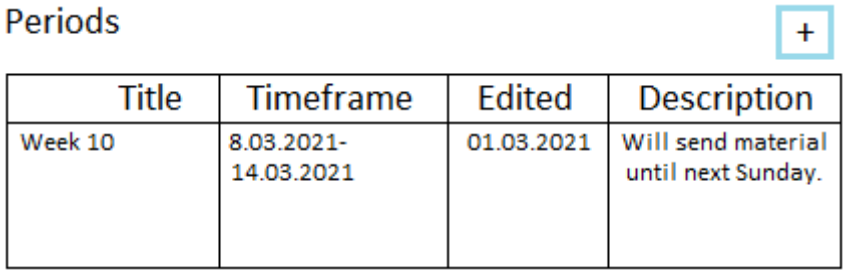

Figure 5. Top layer report view.

Top layer report view opened the top layer maintaining view by clicking + icon. User was able to set the title and the description in the top layer maintaining view (see Fig 6.). Timeframe was set by using the date and the time dropdowns in the top of the view. Day specific timeframe was set in the time dropdowns.

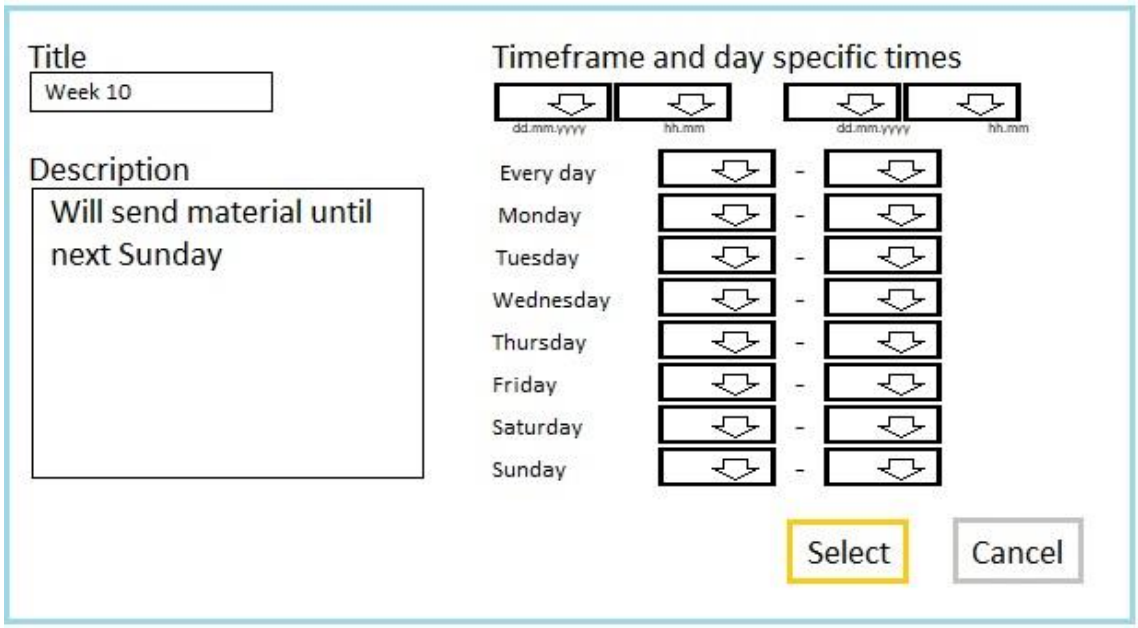

Figure 6. Top layer maintaining view.

After user set the wanted criteria for the selected period the process continued by clicking select button, which opened the material adding view. Material adding view consisted of the two list views (see Fig 7.). The left side list view contained file information and files that were added in the list view to pass the files in the selected targets. The second list

view contained tree view in which the user was able to select the targets where the files will be transmitted.

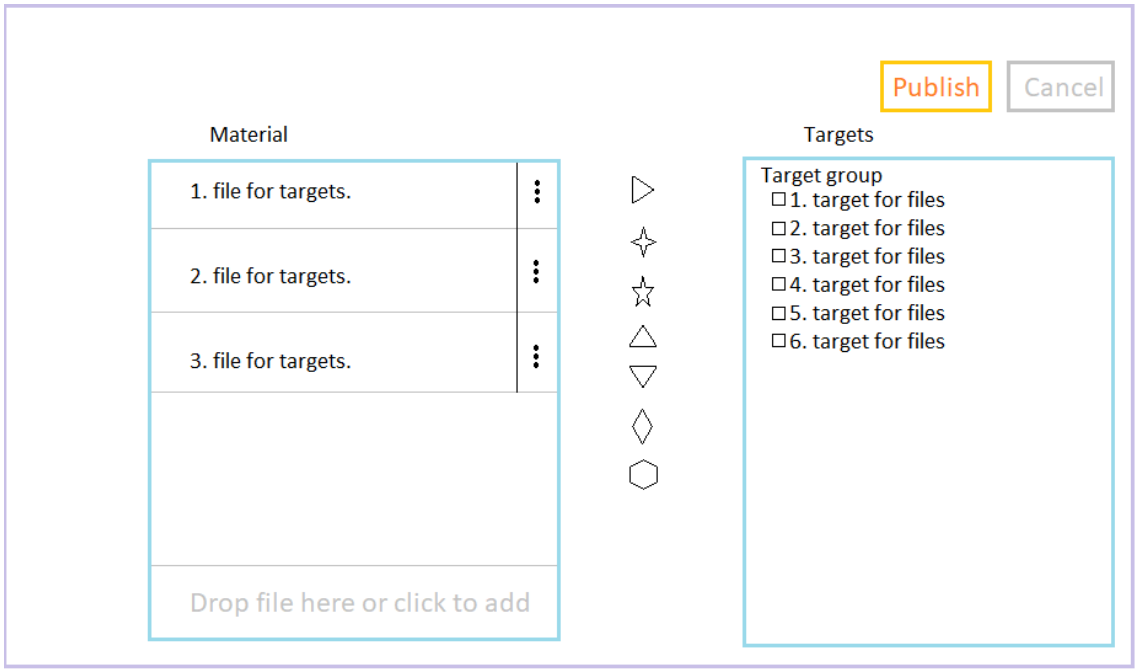

Figure 7. Material adding view.

User may choose the targets and add the files to the first list view. Clicking area to add file opens the file upload view (see Fig 8.). The publish action button was clicked to trigger the action for files to be published in the targets according to the signed schedule after the user had added the files.

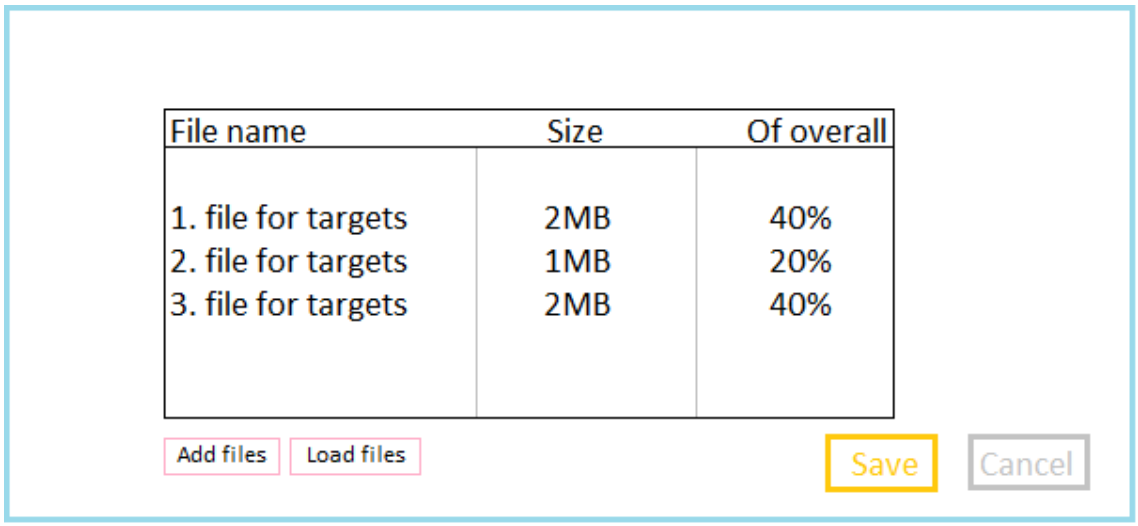

Figure 8. File upload view.

File adding happened through the add files button in the file upload view and clicking the button opened the file selector of the OS in the user's computer. After selecting the files, user loaded the files by clicking the load files to save them in the list by clicking save. Saved files appeared on the material list and the user selected the targets for publish from

the target list. When publish was selected, the material was sent to the targets with their schedules.

5.1.2 List editing

The first model opened edit list view from the home view by clicking the “Edit list” button (see Fig 4.). Edit list view showed two list views with the action buttons (see Fig 9.). List material contained the files in use for the current moment and it was possible to add available files in the list material.

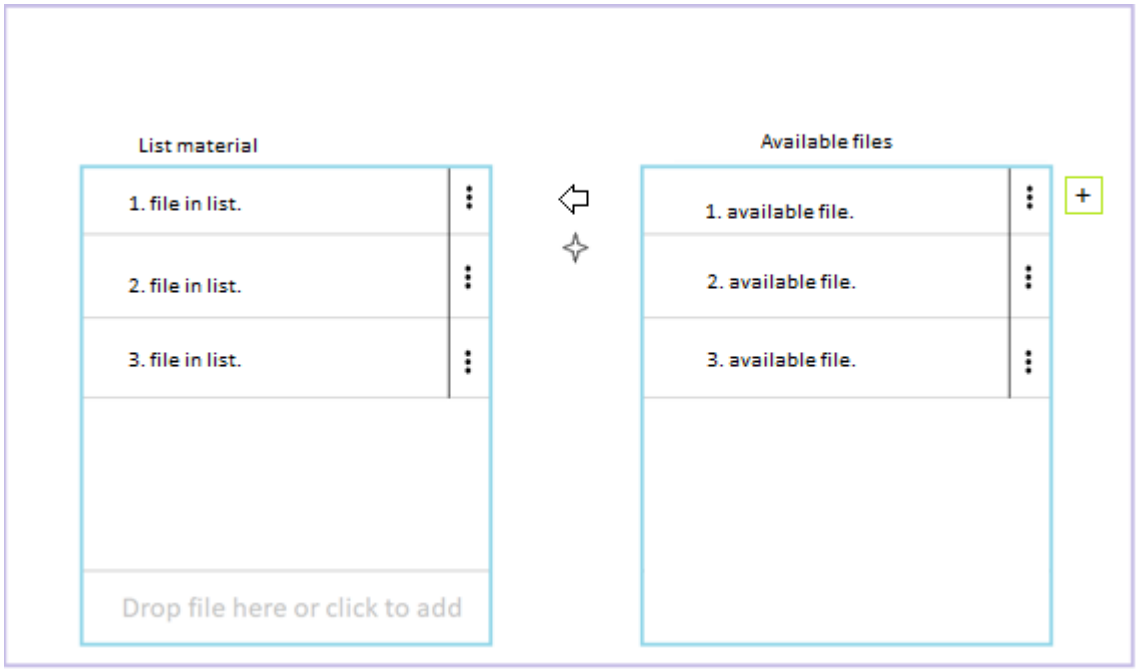

Figure 9. Edit list view.

If user wanted to add a new file in the list material, the first step was to click the plus icon that opened configuration view to set usage preferences to the new file (see Fig 10.). The basic time in use was set to have a time the file usually is used, but the minimum and maximum times were configured to limit other files in use not to consume time for new file that is added.

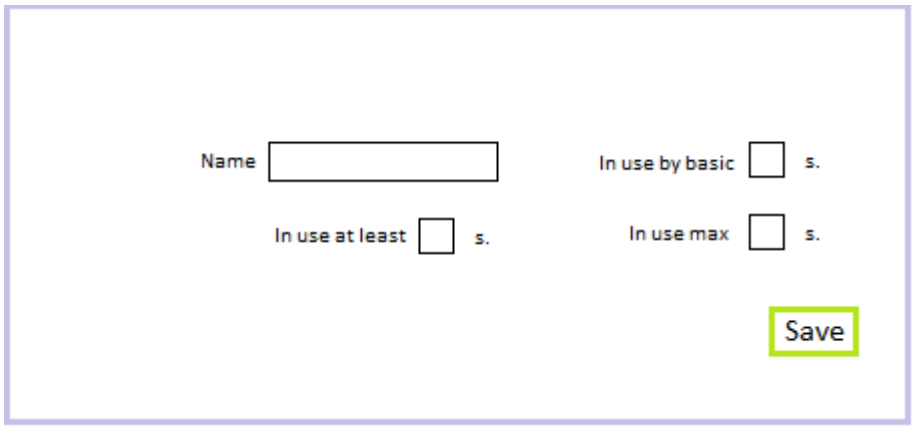

Figure 10. File configuration view.

File adding view was opened when configurations were saved (see Fig 11.). User was able to select a new file for the dropdown list that contained the existing folder structure of the application. Moreover, user was able to add a file from the file system by clicking the diamond icon. The diamond icon opened OS related file addition view.

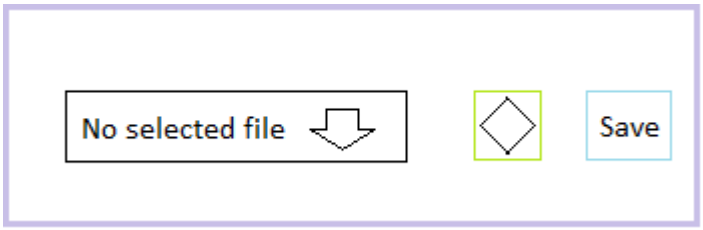

Figure 11. File adding view.

When the user saved the file in file adding view, the file appeared in available the files view (see Fig 9.). The final steps were to move active the file from the available files with arrow to the list material and publish the file by saving the state of the list material by clicking the four-point star.

5.1.3 Target management

User was able to follow the connection state and the file usage state through GUI. The target management page was opened by clicking Targets (see Fig 4.). In target management view there were visible the targets and their states (see Fig 12.).

Figure 12. Target management.

Main use case of the target management was to monitor if targets are operative, inoperative, or not active. Operative cases implicate successful file handling, but other two states will cause need of maintaining the connections and the targets. The Layers tree view was used to filter the shown target groups.

5.1.4 Interview results and model overall

The questions were set in a way that it was possible to follow the main use cases of the different users. Usage context was familiar for the test users as the application had been operational before. Using begun by login to the application, but there was differentiation in regularly used features. The main usages were top layer editing, list view editing and target monitoring.

The question 'How do you use these features' lead in a demonstration about the different features and it was possible to construct a raw drawing usability models about the environment and to point out the improvement possibilities. The usability model was described to be logical when the application was used on daily basis, but for a new user it was seen hard to learn how to use the application. Admin mode feature was hard to find for end users. Adding a file and using the list view were seen complex processes. Test users would not change the top layer editing or the list view editing in general level. Adding new files was seen as a functionality that was really needed and erasing that functionality would have been a flaw for the system. The research gained a slight interest from one of the test user's and the research structure was explained in a deeper level. The found research methods were explained and the purpose behind the usage of chosen methods for this research was reasoned.

There were found six improvement possibilities that had extensive changes in the usability model level. There was a need for button clarification that was seen as a model wide change. In the interview it appeared that the icons alone were not informative enough. This may come up with a hypothesis: *Using only icons to operate with application is not enough to gain good usability.* Adding material in list view was a complex process that had a need for clearer appearance.

Using the list view by side of the tree view was another complex structure. One observation was that the list view was leaving empty slots available, but the empty slots were required in other tree view level and therefore it was reasonable way to describe that the slot was reserved. Another observation was that when user add a date, there was no possibility to double click date to be accepted. Usage of the list view was described as a rare usage feature. This raised up a question that is the rare usage caused by the complexity of the feature.

The findings were raising up the need of usability improvement for the usability model. The model was quite complex and the resourced time span narrow. Fortunately, it was possible to cut the model in smaller pieces.

5.2 TAP in the Second Sprint

Targets were the first improvement scope for the usability model. The main page icons were replaced with guiding 'Targets' text. The model contained new search functionality to locate exact target in the list. New functionality to add the new target was added in the scope.

The first task was to navigate and show where are all the 28 targets available in the Demo target group 1. Users found word Targets appearing in several locations but used the Targets link inside the target group name (See Fig 13.).

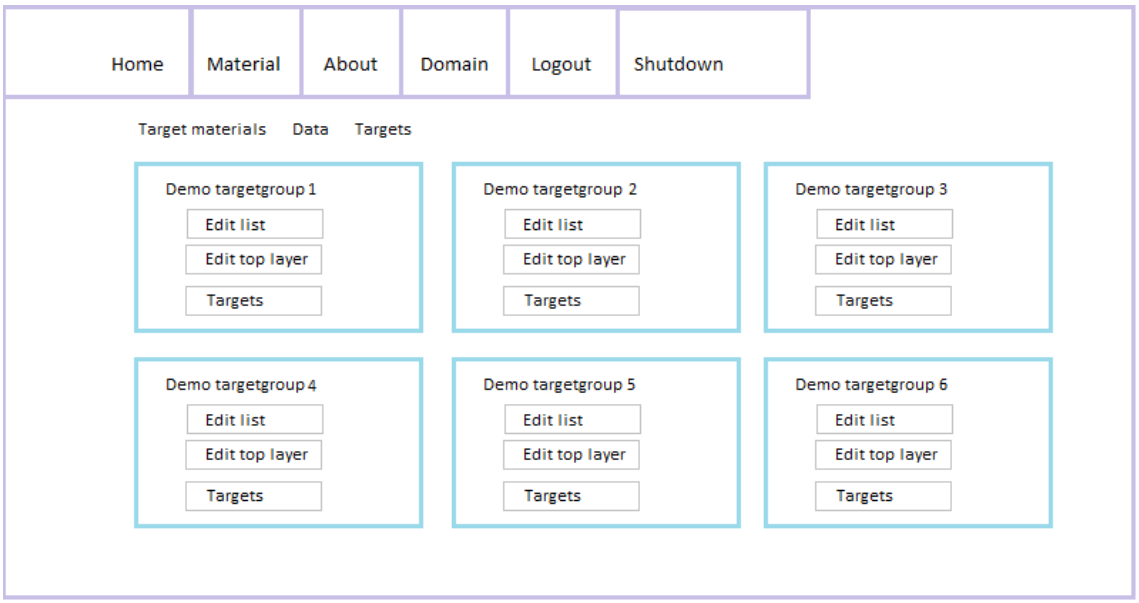

Figure 13. Main view of the second model.

Test users navigated successfully in the target management view. Finding Target 6 was executed with scrolling of the scroll view. There was search tool to find the target, but the test users missed the appearance of the tool (See Fig 14.).

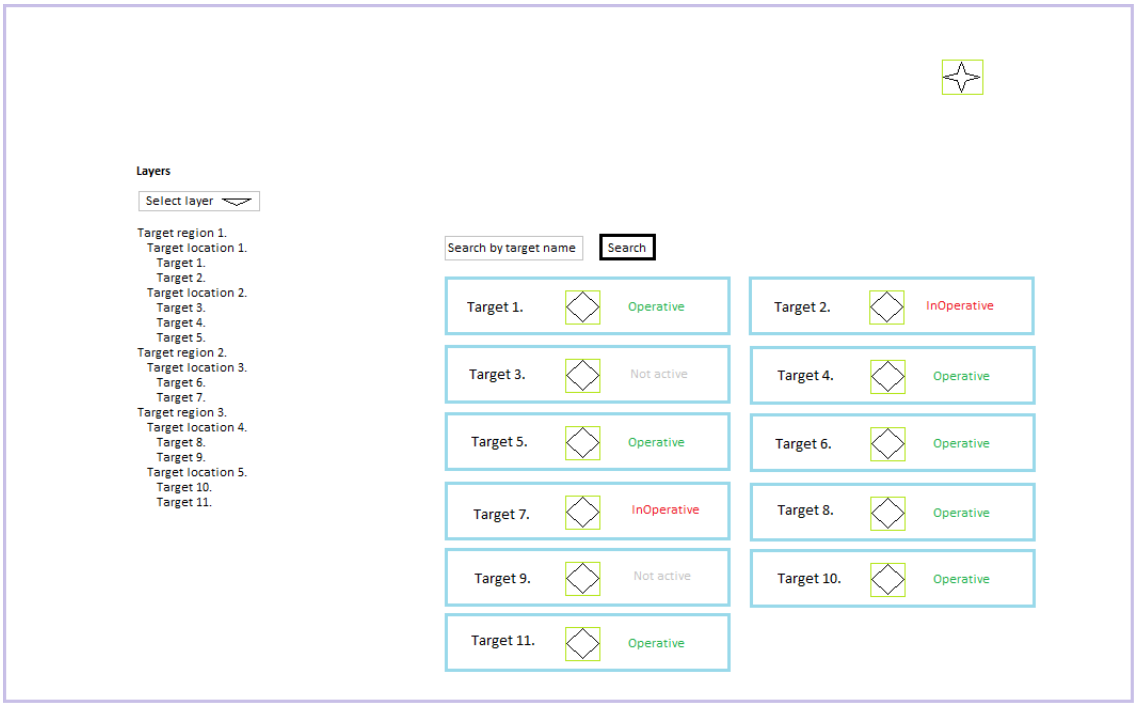

Figure 14. Target management in the second model.

Third task to show the tree view and the targets of the test layer Target location 3 was possible to do in the target management view. Users found the tree view from the left side of the view, but the dropdown of the layer selection was missed feature.

Fourth task was to activate the target with a name Demo target in the Target location. User needed to activate the admin mode from the four-point star in the top right corner of GUI (See Fig 15.). The admin mode was not found by the test users without guidance.

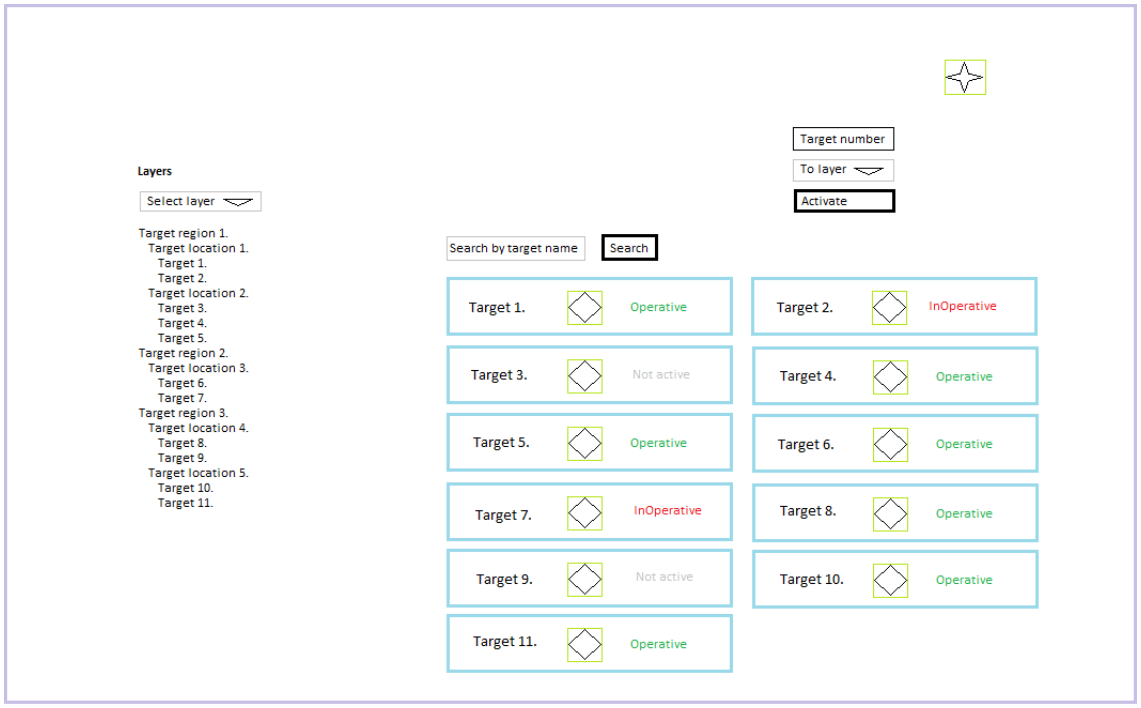

Figure 15. Target management admin mode activated.

Admin mode activation revealed open text field to write target number and dropdown to select what will be the target's destination. Next step was to activate the new target by clicking the activation button. Activation navigated user to the new target adding view where the user was able to set timeframes for the new target (See Fig 16.).

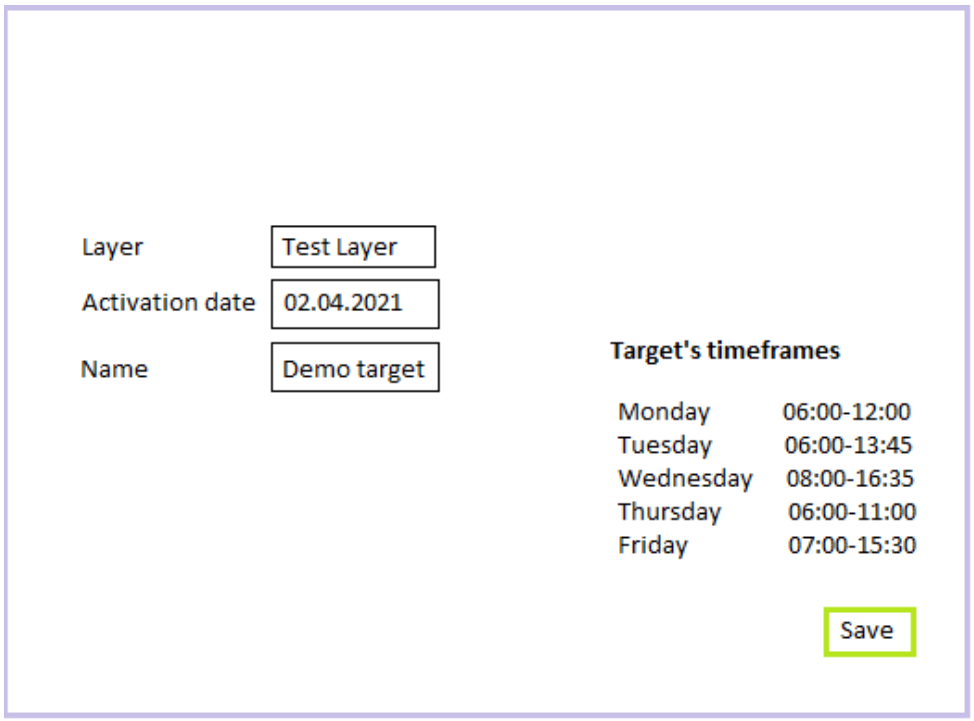

Figure 16. New target adding view.

Finally, test users saved the new target and returned in the target management view. Users were still able to check the new Demo target settings through diamond button in the target management view. There were found six possible improvements from the model after the second sprint.

The improvement targets required smaller changes at this time. Three of the six improvement spots were about the target view and three were more general observations. Admin feature was a feature that was hard to find, and the feature was using only an icon. There was a new search feature to find a target to observe, but that was not used by the test users. Another new feature of the usability model was a dropdown to filter the tree view. It was described that the dropdown menu was hard to find. Tree view's handling was a bit unclear because the current location in the tree view was not highlighted. In the home view Targets text appeared several times and was described as a confusing appearance. The appearance of the application on the screen was in quite small dimensions.

5.3 SUS in the Third Sprint

Core finding from science point of view was that SUS and TAP can be used simultaneously. TAP makes the test user to think about the applications functionalities and it might even brighten the evaluation criteria for SUS. SUS form is in any case filled after the tasks and requires the test user to think about the application after using it when TAP is done simultaneously with the execution of the tasks. This can be formulated as a hypothesis: *SUS and TAP can be used in same test session.*

5.3.1 Proceeding with the model.

Test users navigated in the Edit List view from Main view by clicking Edit List (See Fig 13.). Edit list view contained dropdown selection for tree view in the top left corner (See Fig 17.). Moreover, adding new file was possible from the plus icon in the right side of the view. Available files contained Search file functionality that was hard to find for the test users.

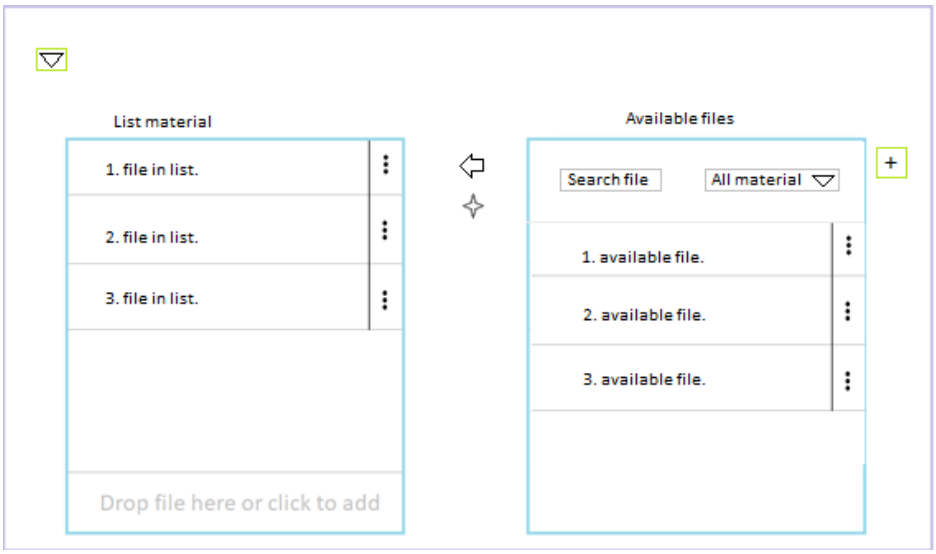

Figure 17. Edit list view in SUS.

The tree view was a missed functionality in three of the four test sessions, but tree view was found as a good addition when the existence was notified for the test users. Tree view allowed user to control which layer's list material is shown and modified. Users were able to open the file editing view by clicking the three dots of the available files list.

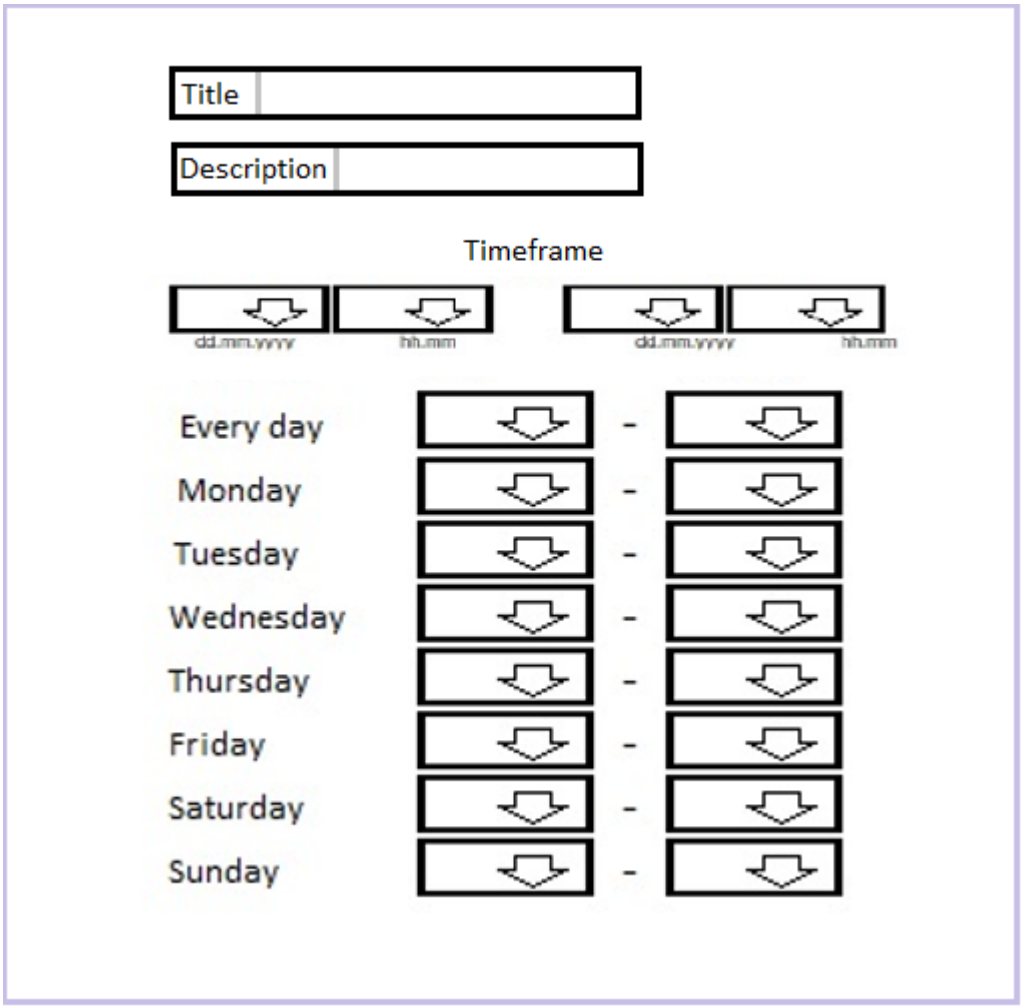

Figure 18. File editing view.

File editing view contained basic operations to manage the file's appearance (See Fig 18.). Test users described file editing view to contain the comprehensive editing material. Scheduling of the file was maintained. It was possible to change the title and the description of the file.

5.3.2 SUS results.

The final number of the participants remain in four. Thesis' target was to have four to ten participants and therefore the participant amount was fulfilled. For quantitative research four participants would be too small number, but this thesis is a case study focusing more on qualitative perspective.

As an average the test users strongly agreed that the application could be used daily (See Table 4.). In this case the user's need for file controlling has been seen through. SUS scale

reveals a flaw in easy to learn aspect. Test users did not find the application easy to learn in a short notice and that can be seen from the average score of 2,5 points for the claim number seven. One slight flaw kind of appearance was with the claim number two with average score of 3,75.

Table 4. Average SUS answers.

	SUS Claim.	Average answer in scale 1-5 where 1 is strongly agree and 5 strongly disagree.
1	From my point of view, it would be nice to use this application on daily basis.	1
2	From my point of view the application felt too complex.	3.75
3	From my point of view, it was effortless to use the application.	1.75
4	From my point of view, it felt like there is a need to contact technical assistant to be able to use the application.	4.25
5	From my point of view, it looked like that the integration of the functionalities was done in a good level.	1.75
6	From my point of view the application was too contradictory.	4
7	It is possible for me to see most of the people learning to use this application in short notice.	2.5
8	It was awkward for me to use this application.	4.5
9	It was self-assured for me to use this application.	2
10	There was a need for me to understand various factors to become fluent user of the application.	4

Overall, the average results present the application usability to be good. Test users did not see the using of the application requiring lots of effort. The implementation of the functionalities was seen as a good level of the realization. It was seen that there is no need for technical assistant to the use of the application and the application is not too contradictory. Using of the application did not feel awkward in the big picture and there was no need to learn various factors for the fluent usage of the application.

5.3.3 SUS score.

Affairs (2013) calculated the SUS score in a scale from zero to hundred be changing the number to give out the scale from zero to forty. The new scale was multiplied by 2.5 to get scale from zero to hundred, but the main thing to remember is that the points are not percentages. 1 was taken off from even indexed claims and odd indexed claims value was taken off from 5 to generate scale from zero to forty (See Table 5.). Sum of those values was 31.5 and multiplied with 2.5 the SUS score resulted to be 78.75.

Table 5. Calculation process.

1	4
3.75	2.75
1.75	3.25
4.25	3.25
1.75	3.25
4	3
2.5	2.5
4.5	3.5
2	3
4	3
Total	31.5
Out of 100	78.75

The trend of the points is over 3 with two exceptions in the ten SUS claims. Total of 31.5 out of 40 is more than three quarters of the points. Those facts already give an impression that the score will exceed 75 points.

Affairs (2013) pointed out that 68 out of 100 points is the average result of the usage. The score 78.75 is well above the average and reaches the laudable appearance. The usability model has succeeded in general level.

6. Discussion

This chapter interacts with all that was done in this case study. Literature review is explained, and different authors views are discussed. Selected research method is reasoned. Hypotheses and results with the literature are summed and the research questions are answered.

6.1 About Literature Review

Material for this research was explored from Scopus article data base search engine. Scopus database was researched with the methods learned from Emerging Trends in Software Engineering course by Mäntylä (2018). The research string was:

TITLE-ABS-KEY(\"Usability Testing\") AND ALL(\"agile\" OR \"continuous deployment\" OR \"heuristic analysis\" OR \"small enterprise\" OR \"small business\" OR \"small company\") AND (SUBJAREA('COMP') OR SUBJAREA('ENGI'))

The overall research resulted in 123 articles. The selection of the different articles for the purposes of this thesis was done by using article selection criteria from course by Ahmad (2018). In the first inclusion or exclusion decision round the articles were selected by the title of the article. Article was excluded from the pool, if any of the research questions of this research were not in the same topic with the title of the found article. In the second inclusion or exclusion decision round the abstracts from the first round's included articles were read.

The second inclusion round lead in a situation where 7 articles were selected as main articles for this research because there was a good answer quality for the research questions. From all the articles that passed the criteria, there were 11 articles that were not available to be read. There were found few more articles around the topic with suggestions made by the fellow scientists.

In usability testing context, Garousi's (2010) research provides the history point of view about the testing in Canada and that is a way to see how things have been handled. Drew et al. (2018) wrote a continuity about that the subjective user perception may have a negative effect on the results because one individual opinion may pull the attention of the researcher away from the essential features. Misunderstanding of the items can lead in misevaluation of the features. The missed improvements can be caused by the progress of the walkthrough because the human mind usually remembers the last things that has been on the mind. A questionnaire set does not provide a proof that the correct features have been evaluated.

For Hodgetts (2006) The general form is not a problem if we try to see that research as a case study. It is good that the article criticizes itself. Most of the text in the article is written from practical experiences of the author of the text. The cause is the lack of the scientific references and therefore the quality of the article is low. The text is not written in scientific form. After all, the Jesse James Garret reference was qualified for the science and is useful for the Hodgetts' research.

The references of the Eck et al. (2018) article do not offer straight path to the source of the paraphrasing. In the article of Meszaros & Aston (2006) The research work is done mostly in practical level and the researchers are using only 5 references in their study. The scientific reliability is low, but the practical experiences are valuable to show how usability testing can be done. In case of Patton (2008), the article is two pages and does not contain scientific references at all. RITE test method is still interesting practical possibility. In the article by Liu et al. (2012), it is partially unclear whose words are used in the text.

The domain expertise and IT background of the user in the article of Bandi & Heeler (2013) could be argued, if the case is that the end users of the software will not be in the IT expertise area. The theory in their research is based around 11 sources which can be enough for short scientific article to have a scientific base. The article itself is only 8 pages and for this reason the explanations could be more comprehensive.

Alomar et al. (2016) had created an article where the text and table results differ from each other and by that way the reliability of the results is a bit questionable. There were 14 references used in the article and it should be scientific enough. There are used various methods to define the usability of the tools under research. From the end user point of view the familiarity with the tools and the software development solves the expected success level during the research.

For Alomar et al. (2016) Task completion is a metric to define how likely the application is successfully used. Number of mouse clicks defines how long the path was to achieve the wanted functionality. SUS operates as the end user point of view scale to define how usable the application was. The tool usage satisfaction level can be used to construct an image about that what is the after all feeling of the users. Observing of the usability problems is a way to find the problems from the developer point of view. The heuristic evaluation rates clear statistics about the tools. There could be at least two more metrics suggested for their study: Time On View (TOV) would be measuring the time spent in one view during the task. View Visits (VV) would be measuring how many times user needs to visit one view to use the application properly. This will give us a hypothesis: *TOV and VV can be used as new heuristic variables in the heuristic analysis.*

In the article of Lizano, Sandoval & Stage (2014) The concept of RS is not explained. The lack of explanation prevents reader to understand the suggested methodology of the integrating usability testing in Scrum. The possibility for the smaller iterative usability evaluations might work. From the article of Barra et al. (2019), it is possible to see that one common aspect in the research area is that the end-users should be taking part to the development process in the early phase. In case of the article by Zein (2016), the definition of PACMAD looks like it would belong in the definition of UCD.

From the article of Bordac & Rainwater (2008) the statement of one user group giving enough input to gain better usability supports small amount of the usability test end user representatives in the practical part of this thesis. Because the usability is about the quality, there is room to move around to find the best way for application to operate. The usability crew in Bordac & Rainwater's research was not able to solve the access to the databases. Reason for this might be that accessing a database is not usability testing requirement.

Research of Gibson and colleagues in years 2016 and 2019 have same conclusions about the applicability of the traditional usability testing methods with the people who have dementia. Research in years 2016 and 2019 came up with the possibilities to improve the usability testing methods for the purposes of the dementia context. Overall, Gibson and colleagues have focused on the application development for three years. As it was stated that the methods could be improved, it would provide the reasons to improve or find the methods for corresponding purposes.

To give more perspective in the article of Hovde (2015) According to the other articles, heuristics are included in the concept of the usability testing. Student feedback could be the end user feedback in the traditional usability testing and there are analytic tools to describe the usability metrics. Therefore, Hovde (2015) could state that the technical communication courses could use usability testing and it covers the whole palette of tools in lectures. The need of the good user experience and the usability in faculty of the technical communications is justified when the faculty is in the professional area of the context.

Hussain, Slany & Holzinger (2009) are stating that the integrating usability with the user focused methods in agile has been widely researched. Nielsen (2012) is disagreeing by stating that only few studies exist about the implementation between the agile software development and the usability testing. These authors' views about the current AUCDI situation are the opposite from each other and the timeline makes disagreeing even more suspicious. In the research of this thesis there were not found various AUCDI type articles with the search string used in the Scopus database. The number of articles found in the Scopus database search are supporting Nielsen's point of view.

Tortor et al. (2019) made it possible to make a hypothesis "After every agile iteration it is possible to organize usability testing." From this material one can expand testable hypothesis: *After every agile iteration it is possible to organize the usability testing with different usability testing methods.*

6.2 Selected Research Method

This section under Discussion was formed to reason the selected interview, TAP, and SUS usability testing methods. First, the top term of the usability testing method with the triangulation is justified. Then the usability testing methods are reasoned one by one in their own sections.

6.2.1 Case Study with Triangulation

Usability testing in general was new topic for the researcher. This made a need to learn about the research topic itself. Research method was selected and created for thesis because it was possible to make a connection between the action research, Agile, and different usability testing methods.

Nielsen and Madsen (2012) gave one more reason to choose usability testing in practical environment as a topic when they wrote about the need of research to combine the Agile and the usability testing methods. In the literature review phase, the idea about the research method grew in its final form when there were found several different types of

the usability testing methods. The chosen method is one combination that can be made under this topic.

The reason to start with this exact method is partially random, but when different usability testing methods were explored the combination felt like a good package to evolve during Agile proceeding. The environment and the software development process brought a hypothesis for choosing usability testing methods during process: *The further the software development proceeds the more reasonable it is to use quantitative usability testing methods.*

Quantitative data hypothesis has been based on increasing the complexity of the application under development. In a complex application, there are lots of options the end user can use, and the end user will spend more time with the application. This simply increases the possibility to collect more quantitative data, which increases the reliability of the quantitative results.

Case study type of the research as a top concept was chosen because in a case study there is room to operate between different kinds of research methods. That freedom provided the possibility to combine the Agile and the usability testing with qualitative and small quantitative approach. Moreover, an action research had a clear connection to the Agile sprints to define the change inside the research itself. The action research was also a match with several existing usability testing methods. It is possible to create a hypothesis for using the research method of this thesis: *Several usability testing methods can be used in Agile software development when sprints operate as an action research's change point.*

6.2.2 Qualitative Beginning

Interview was used as a starter for the research itself. As Myers (2013) explained, in a case study the essential role is to explain *how* and *why* something happens. Qualitative research method offers interview as a tool to describe the application under the usability testing.

The selection to begin with the interview was justified around the hypothesis about as follows: when software development proceeds there is a greater need for quantitative data. Another reason to use the interview and the explanations on *how* and *why* questions is to gain understanding about the operation of the application under development.

When looking at Alomar et al. (2016) writing that usability is about that how end user is using the application, it is possible to reason the usage of an interview to explain the first usability model. As a support for this reason, Bordac & Rainwater (2008) in their case study were researching *how* end users use an application and, in their study, interview was brought up several times to research usability. At this point, we can create a hypothesis: *Interview can be used as a usability testing research method in case study action research using AUCDI.*

6.2.3 TAP Continuing Qualitative Research

For the second sprint it was reasonable to see the usability testing in clear action with the usability model itself. After literature reviewing about TAP there was a tool to research

the usability in action between the test user and the application itself. TAP was described as one of the most traditional usability testing methods by Gibson et al. (2016).

May (2019) had found a connection between LP videos and TAP. LP videos were used to explain for the test users how they should act during the proceeding with the given tasks. May (2019) had a clear vision about the ideal test condition and therefore usage of the TAP was possible to be organized. Hypothesis for TAP usage can follow the line of hypotheses: *TAP can be used as a usability testing method in case study action research using AUCDI.*

As it was found after the first sprint that there was a clear need to develop the usability model in smaller pieces, it was an advantage to have a triangulation in the action research as a research method. Action research not just allow but requires change during the research. Agile software development is built for constantly changing development process. These two methodologies fulfil each other at least in this case.

6.2.4 SUS Verging on Quantitative Research

Third sprint was taking a turn closer to the statistical results. SUS was not seen as a clear quantitative usability testing method, but the SUS is already offering numerical values about the usability. Using the SUS type usability testing method was almost forced for the third sprint when trying to follow the hypothesis about the software development proceeding further.

The literature review brought up that Drew et al. (2018) gave a description about SUS in general. Alomar et al. (2016) were using SUS to define the general usability of the application. Also, Liu et al. (2012) were using SUS in their research. In the concept of Gibson et al. (2016, 2020) the SUS was found unreliable, but for understandable reason when test users were living with dementia. These authors used the SUS in their research about usability and this proves that the SUS is used as a usability testing method in the professional manner.

Results of this case study's implementation of the SUS fulfil the use Alomar et al. (2016) stated to be showing the usability of the application in general level. The success of the usability model made during the process is clear with calculated SUS score and the average results pointed out that the application could perform better if it would be easier to learn use cases. Reliability of the results is tolerable, but increased number of test users would equalize the sharp quarter type results in average scores.

Simultaneous usage of TAP and SUS in Agile software development was not brought up as a statement in the literature reviewed for this thesis. The practical implementation of this thesis showed that the simultaneous usage can be done. This is notable finding for further studies in AUCDI.

In the case of third sprint the functionality base of the application should already be quite known as an environment. This may lead in a situation where developer cannot see the quality of the usability or possible improvement points of the application. Quantitative type of data can give the straight facts about the status of the current model at this point. Hypothesis for SUS follows the type with TAP hypothesis: *SUS can be used as a usability testing method in case study action research with qualitative emphasis using AUCDI.*

6.3 Summing Hypotheses and Results with Literature

There were found nine hypotheses during the research process and eight of the hypotheses were tested (See Table 6.). New heuristic analysis variables will be left for further research to evaluate as usability testing metrics. It was surprising to find nine hypotheses during the research, but the findings confirm usefulness of a case study as a research tool.

Table 6. Collected hypotheses.

Hypothesis	**Operative area of the hypothesis**	**Test status of the hypothesis**
After every agile iteration it is possible to organize usability testing with different usability testing methods.	Research method from literature.	Was tested and can be experimentally tested again with similar type of research structure.
The further the software development proceeds the more reasonable it is to use quantitative usability testing methods.	Usability testing method with relation to the research method.	Was tested and can be experimentally tested again with similar type of research structure.
Several usability testing methods can be used in Agile software development when sprints operate as an action research's change point.	Usability testing method with relation to the research method.	Was tested and can be experimentally tested again with similar type of research structure.
Interview can be used as a usability testing research method in case study action research using AUCDI.	Usability testing method.	Was tested and can be experimentally tested again with similar type of research structure.
TAP can be used as a usability testing method in case study action research using AUCDI.	Usability testing method.	Was tested and can be experimentally tested again with similar type of research structure.
SUS can be used as a usability testing method in case study action research with qualitative emphasis using AUCDI.	Usability testing method.	Was tested and can be experimentally tested again with similar type of research structure.
TOV and VV can be used as new heuristic variables in heuristic analysis.	Heuristic analysis variables.	No resources to test in this study. For further research purposes. Experimental testing of the hypothesis is possible by attaching

		variables in a heuristic analysis.
Using only icons to operate with application is not enough to gain good usability.	Practical solution in this exact thesis.	Was tested and can be experimentally tested again by using icons only to navigate in an application.
SUS and TAP can be used in same test session.	Practical solution in this exact thesis.	Was tested successfully and can be experimentally tested again by organizing usability testing environment where test user executes tasks.

All the tested hypotheses were successful. It requires more testing of the found hypotheses to conduct a theory. Literature supported the finding of the hypotheses. In case of the new heuristic variables, they were found by reading the existing heuristic variables and perceiving the operational environment of any application. Using different usability testing methods after every Agile iteration was conducted from literature where researchers had used one of the usability testing method after iterations. Rest of the hypotheses were found by practical implementation of this case study.

6.3.1 How should the usability testing be done in agile development context?

The literature revealed that the answer is not a one correct explanation for this question. Bordac & Rainwater (2008) used a trial-error-model when Nielsen & Madsen (2012) explained that the only way to find any solutions is to try different types of the usability testing scenarios to create an operative AUCDI. This case study proposes one way to do usability testing in Agile software development. There are many other ways, but literature about different methods was hard to find.

This case study proposed a way to do the usability testing with a new research type of approach in AUCDI concept. Moreover, this research method type is adaptive for various scenarios in this context. Usability testing can be done by following the research method of this study by collecting different types of the usability testing method sets in use.

One suggestion to conduct usability testing in Agile environment is that if SUS or TAP is used, both can be used in the same session. This gives access in numerical data and qualitative descriptions about the application at the same time. The approach of this case study executed usability testing sessions at the end of the sprints when usability models were done during the sprints.

One suggested next step after this type of usability testing would be to use the heuristics. Good starting point for that kind of study would be to follow Alomar et al. (2016) with

their heuristics and adding the suggested new heuristic variables in the research. Another statistical approach would be to use Likert scale, that was used by Hussain, Slany and Holzinger (2009).

6.3.2 What are the benefits reached with specific usability testing methods in agile development context?

Semi structured interview offered a benefit of having more like a conversation type moment with the test user. In this way it is possible to find new ways to approach the usability of the target application by free type of the conversation. When interview can be recorded, there is more time to focus on the conversation than taking down the notes.

Literature about TAP was more about that how this usability method should be used. The descriptions about TAP are still offering on the reached main advantages. One benefit is that TAP allows test user to speak out the actioning when using an application. This way the reactions about the application are more instant than in the written mode. Another advantage is that the researcher can focus on the observing when the test user is the one who is speaking. Thirdly, there is no need for the test user to write notes if voice recording is allowed.

Drew et al. (2018) pointed out various advantages and disadvantages in usability testing overall. In case of SUS there was possibility to misunderstand the items in both end user´s and researcher´s end (Drew et al., 2018, p. 358). Then again Drew et al. (2018) wrote about that easiness to respond from the participant's point of view, the quick questionnaire, free, and managing friendly are advantages of the SUS (Drew et al., 2018, p. 357). On the other hand, the main advantage of SUS was the setup of the questions. SUS offered reliable quality assessment with low quantity need of the test users. Benefits in this case study were that the SUS pointed out a flaw in easy to learn aspect of the application and gave high SUS score to bring confident to the usability of application's features.

Liu et al. (2012) used crowdsourcing as a tool to success with usability testing. Crowdsourcing offered a cheap way to gain big quantity of test users in the usability testing. Web development operates as a good basis for crowdsourcing because then the test users are not tied to be in the same place with the device.

Gibson et al. (2016, 2019) used REC as a usability testing method. In their context REC was found more disturbing the test users than being a helpful tool. REC may offer material about the actual system user interaction, but the context must be suitable to use REC. Another usability testing method used by Gibson et al. (2016) was TCR. TCR is a metric giving an advantage by indicating how successfully test users were able do given tasks. Third tool used by Gibson et al. (2016) was SEQ which was used before and after the tasks. SEQ may offer SUS type scale about the usability. Therefore, advantages are nearly similar with SUS because SEQ can give numerical values about the usability itself.

One good use case for LP videos is to operate as an example for the test user. LP videos can show how TAP should be done. Differing application context will change the description topics, but the idea of talking when progressing with the application is the catch.

Observing end user is giving an opportunity for developers to see, how the user is using the application. Taking notes about users' actions while observing offers various possibilities to see if end user is using application in a way the developer has not planned it to be used. Literature material revealed that the widest usage of the usability testing method was for observing the end user. Meszaros & Aston (2006), Bordac & Rainwater (2008), Garousi (2010), Liu et al. (2012), Wetzlinger et al. (2014), Potts, Nguyen & Turner (2016), (Osorio, Aristizábal & Zuluaga (2016), Alomar et al. (2016), Gibson et al. (2016), Germanakos (2018), Eck et al. (2018), May (2019), Tortor et al. (2019) and, Barra et al. (2019) mentioned observing of the end user as a usability method. Barra et al. (2019) outlined that the purpose of the PlatoS was to reduce the need of observers as a person to be observing when usability testing is done. The usage level of observing as a usability testing tool might indicate that observing is an efficient way to find out flaws and misleading functionalities in the usability.

6.3.3 How could the usability testing process be improved in agile development context?

Literature does not give a simple answer to this question. The first step to improve the usability testing is to take at least one usability testing method in use. This thesis is offering a set of usability testing methods and there might be even more methods to use. The collection of different usability testing methods gives a possibility to compare the methods with context of the application under test.

Any other usability testing method set with AUCDI might have better performance than the one used for this research. There might be various tools to improve usability testing and it looks like the answer is the case specific selection.

One central thing was common in the literature. In usability testing and software development processes, the end user should be taken as a part of the practice. It was done in the implementation of this thesis and resulted in the feedback to prioritize changes in the usability model. Next improvement step for this case study would be the usage of the heuristic analysis.

6.3.4 How could the usability testing be brought in Agile and SCRUM environments?

Few ways of AUCDI were found from literature in this case study. There were used the usability testing during the sprints for small parts of the application under the development in few of the studies. Another use case was to conduct usability testing every time at the end of the sprint.

This case study brought usability testing in Agile SCRUM environment by placing the usability testing at the end of the first sprint. Then usability testing was done at the end of sprints two and three when the usability model parts were refined. Usability model creation is a way to show how adaptive usability testing is even for Agile environment.

Agile and usability testing are separate tools, but it is possible to run both in parallel. This can bring the tools in the same environment as co-operative methods. Separation again can give a possibility to run the usability testing without following Agile proceeding

strictly. Usability model creation is a lot faster tool to see how application should function. Therefore, the usability testing can give information about the application behaviour before programmed solutions are created. This causes efficiency in the development process by reducing the costs when the usability model can inform how the back end should operate.

6.3.5 Implications to theory

The researchers of the usability testing and Agile will benefit of this case study as a collection of different usability testing methods. Reviewing the literature made it possible to gather a comprehensive set of the usability testing tools in one paper. Moreover, this study successfully used the selected usability testing methods in Agile software development process. The described successful usage is a proof of that AUCDI is possible.

The implementation and the findings of this study used interview, TAP, and SUS to define usability of an application. For theory, the implementation and the findings mean that those tools can be used to conduct usability testing in Agile development context. One should take a note about that TAP and SUS can be used in a same test session. Using TAP and SUS in the same test session will give more to evaluate in a sprint.

This study constructed a research method that can be used to test different sets of the usability testing methods. It is notable that the action research enables a researcher to adapt in a change during a sprint in Agile development context. The constructed research method and the possibilities of the action research in Agile development context offer a new approach for AUCDI to evolve.

6.3.6 Implications to practice

The practitioners of the usability testing and Agile will benefit of this case study as a justifying background to unite the usage of both methods. Usability testing practitioners with knowledge about SUS and TAP will benefit of this study because it shows SUS and TAP integrated in Agile development context. Interview is not that common tool used in usability testing. Therefore, this study encourages practitioners to interview the test users of their usability models to gain qualitative understanding about the application from test user point of view.

In more general level, the practitioners of the usability testing and Agile can be encouraged to try more different usability testing methods in Agile development context. Eventually, the trying process can lead in a situation where it is possible to gather a set of applicable usability testing methods for Agile development context. It would be valuable to gather a set of applicable tools to quickly catch how to begin specializing in the field.

The constructed research method is a tool for practitioners to understand one way of using usability testing methods. For more, the tested hypotheses suggest that it is possible to use different usability testing method after every Agile iteration. When constructed research method and different usability testing methods are combined, there is a variety of combinations to be used for usability testing.

6.3.7 Future research

More solutions for AUCDI could be found at least from REC, CrowdSourcing, TCR, and SEQ. Those tools could be examined in similar type of research method setting, but it is possible to construct a new research method for further research purposes. Deeper examination of the appeared 120 articles might offer surprises in a form of different usability testing methods, if the topics or abstracts of the articles did not offer enough material with good relation to the research questions.

Using quantitative research tools for AUCDI is another way to bring more data about that how usability testing tools can succeed. Quantitative research would suggest usage of the quantitative usability testing methods too. Heuristics are one way to gather quantitative data out of the usage of an application. This study suggested new metrics TOV and VV that could be used in heuristic evaluation. Those metrics must be tested in practice before evaluating how they describe the usability. TOV and VV would need a collected data bank to analyze what are the expected values.

There is a lack of the scientific material about usability testing in Agile development context. Science is offering various tools to research any topic and that means there is still a lot to be researched in this context. The hypotheses this study claimed should be tested and evaluated by other researchers. More hypotheses could be found during research of other usability testing methods.

7. Conclusions

This Master's thesis conducted a case study research with triangulation to an action research which utilizes the research tools found from the literature review. Introduction described the operational environment and reasoned shortly the importance of the usability testing. Research questions were also set in the introduction. Literature Review followed Introduction to describe the theoretical environment of the usability testing itself.

The Research Method was gathered from theoretical qualitative research material. SUS, TAP, and UCD were adapted with the use of the qualitative research method to generate a research tool for this case study. The Implementation explained more about the operational environment and described how the usability testing sessions were set.

Findings described what was found in the usability testing sessions. Each sprint's usability testing session findings were divided under their own sub headers. Interview, TAP, and SUS were used as tools to conduct UCD. The UCD was achieved by taking test users from the users of the application.

Discussion interacted with the researched literature, reasoned the selected usability method and usability testing tools, combined Findings and Literature Review to discuss together, and collected found hypotheses to answer the research questions. Finally, Conclusions explained what was done and what was derived from the case study after all.

There were found various tools to conduct usability testing. Results of this exact research followed the usability testing principals with a new approach to conduct AUCDI. Nine new hypotheses were introduced. Successful testing of the hypotheses can encourage the usability testing professionals to the use usability testing methods in Agile software development. This thesis took few steps to evolve usability testing from science point of view.

Literature offered understanding about the different usability testing environments. Finding suitable material for this topic was challenging. Therefore, the number of sources remain low in numbers. The topic needs more scientific research to support practical solutions for the software development.

Practical part of the thesis conducting the usability testing offered possibility to try how the usability testing methods can be implemented in the software development process. Finding out that TAP and SUS can be used in the same usability testing session was very instructive for the future. In future research, the same setup can include SUS in all TAP sessions.

Another surprise was that the action research and Agile can fulfil each other when doing the usability testing. Action research required change and Agile created change. Action research could be recommended as a research tool for further Agile research.

There was not found explanation how the usability testing should be done. Still, there were found several usability testing methods that can be applied. The found usability testing methods may be only few of the existing ones. This case study collected several

testing methods together to make it easier finding suitable testing method for the usability testing purposes.

There were found benefits for various testing methods. The main findings were in the implemented usability testing methods. TAP offers descriptive information how the test user experiences the application. Observing the usage offers possibility to find the usage paths to verify if the application has been understood properly. SUS offers possibility to find out the usability level of the application in general level. Interview gives more room for the test user and interviewer to discuss about the application's behaviour. It is recommended to familiarize with the other usability testing method explained in this case study if one is looking for an appropriate usability testing method for some other case study or the usability testing session.

The simplest way to improve the usability testing process is to use at least one usability testing method. If one is already using usability testing, next step would be to take the end user of the application in the test sessions. Once again, there is no one correct solution to conduct usability testing. Different testing methods can explain one way and decision can be made by thinking how the method fits the goal of the usability testing.

Usability testing was brought in Agile environment by executing the usability tests after every sprint in this case study. This was a successful way to conduct usability testing and action research operated as a synchronic pair with Agile. Usability model creation can work faster than development in a sprint.

References

Affairs, A. S. for P. (2013, September 6). System Usability Scale (SUS). Usability.gov. https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

Ahmad, M. (2018). *Software Engineering Research.* Retrieved from: University of Oulu. 815663S.

Alomar, N., Almobarak, N., Alkoblan, S., Alhozaimy, S., & Alharbi, S. (2016). Usability Engineering of Agile Software Project Management Tools. *Design, User Experience, and Usability: Design Thinking and Methods Lecture Notes in Computer Science*, 197–208. doi: https://doi.org/10.1007/978-3-319-40409-7_20

Bandi, A., & Heeler, P. (2013). Usability testing: A software engineering perspective. *2013 International Conference on Human Computer Interactions (ICHCI)*. doi: https://doi.org/10.1109/ichci-ieee.2013.6887809

Barra, S., Francese, R., & Risi, M. (2019). Automating Mockup-Based Usability Testing on the Mobile Device. *Green, Pervasive, and Cloud Computing Lecture Notes in Computer Science*, 128–143. doi: https://doi.org/10.1007/978-3-030-19223-5_10

Bordac, S., & Rainwater, J. (2008). User-Centered Design in Practice: The Brown University Experience. *Journal of Web Librarianship*, *2*(2-3), 109–138. doi: https://doi.org/10.1080/19322900802205908

Choi, Y. M. (2015). Utilizing end User Input in Early Product Development. *Procedia Manufacturing*, *3*, 2244–2250. doi: https://doi.org/10.1016/j.promfg.2015.07.368

Drew, M. R., Falcone, B., & Baccus, W. L. (2018). What Does the System Usability Scale (SUS) Measure? *Design, User Experience, and Usability: Theory and Practice Lecture Notes in Computer Science*, 356–366. doi: https://doi.org/10.1007/978-3-319-91797-9_25

Eck, M. L. V., Markslag, E., Sidorova, N., Brosens-Kessels, A., & Wil M. P. Van Der Aalst. (2018). Data-Driven Usability Test Scenario Creation. *Human-Centered Software Engineering Lecture Notes in Computer Science*, 88–108. doi: https://doi.org/10.1007/978-3-030-05909-5_6

Ferre, X., Villalba, E., Julio, H., & Zhu, H. (2017). Extending Mobile App Analytics for Usability Test Logging. *Human-Computer Interaction – INTERACT 2017 Lecture Notes in Computer Science*, 114–131. doi: https://doi.org/10.1007/978-3-319-67687-6_9

Garousi, V., & Varma, T. (2010). A replicated survey of software testing practices in the Canadian province of Alberta: What has changed from 2004 to 2009? *Journal of Systems and Software*, *83*(11), 2251–2262. doi: https://doi.org/10.1016/j.jss.2010.07.012

Germanakos, P., & Fichte, L. (2018). EUREKA: Engineering Usability Research Empirical Knowledge and Artifacts. *Learning and Collaboration Technologies. Design, Development and Technological Innovation Lecture Notes in Computer Science*, 85–103. doi: https://doi.org/10.1007/978-3-319-91743-6_6

Gibson, A., Mccauley, C., Mulvenna, M., Ryan, A., Laird, L., Curran, K., … Bond, R. (2016). Assessing usability testing for people living with dementia. *Proceedings of the 4th Workshop on ICTs for Improving Patients Rehabilitation Research Techniques - REHAB 16*. doi: https://doi.org/10.1145/3051488.3051492

Gibson, A., Mccauley, C., Mulvenna, M. D., Ryan, A., Laird, E. A., Curran, K., … Bond, R. (2019). How Usable Are Usability Tests? Examining the Suitability of Standard Usability Testing Methods for the Assessment of Apps for People Living with Dementia. *New Technologies to Improve Patient Rehabilitation Communications in Computer and Information Science*, 126–143. doi: https://doi.org/10.1007/978-3-030-16785-1_10

Hodgetts, P. (2006). Experiences integrating sophisticated user experience design practices into agile processes. *Agile Development Conference (ADC05)*. doi: https://doi.org/10.1109/adc.2005.24

Hosseini-Khayat, A., Hellmann, T. D., & Maurer, F. (2010). Distributed and Automated Usability Testing of Low-Fidelity Prototypes. *2010 Agile Conference*. doi: https://doi.org/10.1109/agile.2010.11

Hovde, M. R. (2015). Effective user experience in online technical communication courses. *Proceedings of the 33rd Annual International Conference on the Design of Communication - SIGDOC 15*. doi: https://doi.org/10.1145/2775441.2775453

Hussain, Z., Slany, W., & Holzinger, A. (2009). Current State of Agile User-Centered Design: A Survey. *HCI and Usability for e-Inclusion Lecture Notes in Computer Science*, 416–427. doi: https://doi.org/10.1007/978-3-642-10308-7_30

Jarraya, M., & Moussa, F. (2018). Proxy Oriented Approach for Evaluating Usability of a Resilient Life-Critical Interactive Systems. *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*. doi: https://doi.org/10.1109/aina.2018.00075

Larusdottir, M. K., Bjarnadottir, E. R., & Gulliksen, J. (2010). The Focus on Usability in Testing Practices in Industry. *IFIP Advances in Information and Communication Technology Human-Computer Interaction*, 98–109. doi: https://doi.org/10.1007/978-3-642-15231-3_11

Liu, D., Bias, R. G., Lease, M., & Kuipers, R. (2012). Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. doi: https://doi.org/10.1002/meet.14504901100

Lizano, F., Sandoval, M. M., & Stage, J. (2014). Integrating Usability Evaluations into Scrum: A Case Study Based on Remote Synchronous User Testing. *Lecture Notes in Computer Science Human-Computer Interaction. Theories, Methods, and Tools*, 500–509. doi: https://doi.org/10.1007/978-3-319-07233-3_46

Losada, B. (2018). Flexible Requirement Development through User Objectives in an Agile-UCD Hybrid Approach. *Proceedings of the XIX International Conference on Human Computer Interaction - Interacción 2018*. doi: https://doi.org/10.1145/3233824.3233865

May, J. (2019). YouTube Gamers and Think-Aloud Protocols: Introducing Usability Testing. *IEEE Transactions on Professional Communication*, *62*(1), 94–103. doi: https://doi.org/10.1109/tpc.2018.2867130

Meszaros, G., & Aston, J. (2006). Adding Usability Testing to an Agile Project. *Agile 2006 (Agile06)*. doi: https://doi.org/10.1109/agile.2006.5

Myers, M. D. (2013). *Qualitative research in business & management*. Los Angeles, California: SAGE.

Mäntylä, M. (2018). *Emerging Trends in Software Engineering.* Retrieved from University of Oulu. 811600S.

Nielsen, L., & Madsen, S. (2012). The usability experts fear of agility. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction Making Sense Through Design - NordiCHI 12*. doi: https://doi.org/10.1145/2399016.2399057

Osorio, S. P., Aristizabal, L. M., & Zuluaga, C. A. (2016). Development of a command interface based on handheld devices for remotely operated vehicles. *2016 IEEE Colombian Conference on Robotics and Automation (CCRA)*. doi: https://doi.org/10.1109/ccra.2016.7811422

Ovad, T., & Larsen, L. B. (2015). The Prevalence of UX Design in Agile Development Processes in Industry. *2015 Agile Conference*. doi: https://doi.org/10.1109/agile.2015.13

Patton, J. (2008). Getting Software RITE. *IEEE Software*, *25*(3), 20–21. doi: https://doi.org/10.1109/ms.2008.62

Pillalamarri, S. S., Huyett, L. M., & Abdel-Malek, A. (2018). Novel Bluetooth-Enabled Tubeless Insulin Pump: A User Experience Design Approach for a Connected Digital Diabetes Management Platform. *Journal of Diabetes Science and Technology*, *12*(6), 1132–1142. doi: https://doi.org/10.1177/1932296818804802

Potts, L., Nguyen, M.-T., & Turner, H. (2016). Notation System for Rapid Usability Testing. *Proceedings of the 34th ACM International Conference on the Design of Communication - SIGDOC 16*. doi: https://doi.org/10.1145/2987592.2987635

Rahayu, P., Sensuse, D. I., Fitriani, W. R., Nurrohmah, I., Mauliadi, R., & Rochman, H. N. (2016). Applying usability testing to improving Scrum methodology in develop assistant information system. *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. doi: https://doi.org/10.1109/icitsi.2016.7858222

Rajanen, M., & Iivari, N. (2007). Usability Cost-Benefit Analysis: How Usability Became a Curse Word? *Lecture Notes in Computer Science Human-Computer Interaction – INTERACT 2007,* 511-524. doi:10.1007/978-3-540-74800-7_47

Reeves, S. (2019). How UX Practitioners Produce Findings in Usability Testing. *ACM Transactions on Computer-Human Interaction*, *26*(1), 1–38. doi: https://doi.org/10.1145/3299096

Salah, D., Paige, R., & Cairns, P. (2014). A Practitioner Perspective on Integrating Agile and User Centred Design. doi: https://doi.org/10.14236/ewic/hci2014.14Kitchenham, B. (2004). *Procedures for performing systematic reviews.* Technical Report.TK/SE-0401, Keele University, and Technical Report 0400011T.1, NICTA.

Tarkkanen, K., Reijonen, P., Tétard, F., & Harkke, V. (2013). Back to User-Centered Usability Testing. *Lecture Notes in Computer Science Human Factors in Computing and Informatics*, 91–106. doi: https://doi.org/10.1007/978-3-642-39062-3_6

Tarkkanen, K., Harkke, V., & Reijonen, P. (2015). Testing the Unknown – Value of Usability Testing for Complex Professional Systems Development. *Human-Computer Interaction – INTERACT 2015 Lecture Notes in Computer Science*, 300–314. doi: https://doi.org/10.1007/978-3-319-22668-2_24

Tortor, R. N., Dolosa, R. E., Labay, B. J. S., & Hernandez, J. T. (2019). An Online Workflow Management System for Streamlining the Processing of Business Permits in A Local Government Unit. *Proceedings of the 2019 5th International Conference on E-Business and Mobile Commerce - ICEMC 2019*. doi: https://doi.org/10.1145/3332324.3332332Kitchenham, B. (2004). *Procedures for performing systematic reviews.* Technical Report.TK/SE-0401, Keele University, and Technical Report 0400011T.1, NICTA.

Wetzlinger, W., Nedbal, D., Auinger, A., Grossauer, C., Holzmann, C., & Lettner, F. (2014). Mobile Usability Testing Requirements and their Implementation in the Automation Engineering Industry. *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia - MoMM 14*. doi: https://doi.org/10.1145/2684103.2684120

Winter, J., Rönkkö, K., & Hellman, M. (2009). Reporting usability metrics experiences. *2009 ICSE Workshop on Cooperative and Human Aspects on Software Engineering*. doi: https://doi.org/10.1109/chase.2009.5071422

Zein, S., Salleh, N., & Grundy, J. (2016). A systematic mapping study of mobile application testing techniques. *Journal of Systems and Software*, *117*, 334–356. doi: https://doi.org/10.1016/j.jss.2016.03.065

Appendix A. Consent Form to Usability Testing Research

Consent Form

The purpose of this consent form is to permit interviews and data collection for an academic Master´s thesis. The topic of the thesis is considering the implementation of usability testing in agile software development environment. Data collection forms of the research are interviews, observations in testing sessions, voice recordings in Think Aloud Protocol (TAP) session and forms filled in System Usability Scale (SUS) questionnaire. The collected data will be processed to conduct the results for this research. The conducted results are used in studies like Master´s thesis written by Sakari Järvelä. The purpose of the Master´s thesis is to offer various ways of usability testing implementation with agile software development environment. Moreover, scientific manner of this research is to evaluate existing theories and possibly form new ones.

Your support for this thesis is highly appreciated.

It is in my knowledge that signing of this consent form:

- Is a confirmation that I have understood the reason for information collection.
- Is a confirmation that the author of the research has given possibility to ask questions about the research.
- Is a confirmation that the author of the research has given satisfying answers to my questions.
- It is in my knowledge that participating in the research is voluntary and I can stop participating any time I want to.
- It is in my knowledge that I have possibility to ask my data to be destroyed at any moment. It is also in my knowledge that my data will be destroyed when the thesis has been accepted in University of Oulu.
- It is in my knowledge that my personal data is handled as confidential information.

I accept data usage in academic research explained above.

☐ I give a permission to audio record interviews regarding to this project.

☐ I give a permission to observe my actions with application under testing regarding to this project.

☐ I give a permission to collect data from my System Usability Scale form regarding to this project.

☐ I am willing to offer data in the research described above.

_____ _____ _____

Name Date Signature

Researcher contact:

Sakari Järvelä

sakari.jarvela@student.oulu.fi

Appendix B. SUS That Was Used

System Usability Scale (SUS)

	Strongly agree				Strongly disagree
	1	2	3	4	5
From my point of view, it would be nice to use this application on daily basis.					
From my point of view the application felt too complex.					
From my point of view, it was effortless to use the application.					
From my point of view, it felt like there is a need to contact technical assistant to be able to use the application.					
From my point of view, it looked like that the integration of the functionalities was done in a good level.					
From my point of view the application was too contradictory.					
It is possible for me to see most of the people learning to use this application in short notice.					
It was awkward for me to use this application.					
It was self-assured for me to use this application					
There was a need for me to understand various factors to become fluent user of the application.					